인류학 석사학위논문

공공재 게임에서 나타나는 평판 평가의 특성

Characteristics of reputation evaluation

in the public goods game

2013년 2월

서울대학교 대학원

인류학과 진화심리학 전공

이시원

인류학 석사학위논문

공공재 게임에서 나타나는 평판 평가의 특성

Characteristics of reputation evaluation

in the public goods game

지도교수: 박순영

이 논문을 인류학 석사 학위 논문으로 제출함

2012년 12월

서울대학교 대학원

인류학과

이시원

이시원의 석사 학위 논문을 인준함

2012년 12월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

학위논문 원문제공 서비스에 대한 동의서

본인의 학위논문에 대하여 서울대학교가 아래와 같이 학위논문 제공하는 것에 동의합니다.

1. 동의사항

① 본인의 논문을 보존이나 인터넷 등을 통한 온라인 서비스 목적으로 복제할 경우 저작물의 내용을 변경하지 않는 범위 내에서의 복제를 허용합니다.

② 본인의 논문을 디지털화하여 인터넷 등 정보통신망을 통한 논문의 일부 또는 전부의 복제.배포 및 전송 시 무료로 제공하는 것에 동의합니다.

2. 개인(저작자)의 의무

본 논문의 저작권을 타인에게 양도하거나 또는 출판을 허락하는 등 동의 내용을 변경하고자 할 때는 소속대학(원)에 공개의 유보 또는 해지를 즉시 통보하겠습니다.

3. 서울대학교의 의무

① 서울대학교는 본 논문을 외부에 제공할 경우 저작권 보호장치(DRM)를 사용하여야 합니다.

② 서울대학교는 본 논문에 대한 공개의 유보나 해지 신청 시 즉시 처리해야 합니다.

논문 제목 : 공공재 게임에서 나타나는 평판 평가의 특성

학위구분 : **석사 ▣ · 박사 □**
학 과 : 인류학과
학 번 : 2010-22986
연 락 처 :
저 작 자 : 이시원 (인)

제 출 일 : 2013 년 1 월 일

서울대학교총장 귀하

<국문초록>

서울대학교 대학원
인류학과 진화심리학 전공
이시원

인간의 이타적 행동은 그 범위가 친족의 범주를 벗어난다는 점과 현재의 이타적 행동의 수혜자와 미래의 상호작용을 확신할 수 없다는 점에서 독특하다. 이 때 관찰자의 유무에 따라 이타적 행동에 차이가 있음을 통해서 인간의 이타적 행동은 무조건적인 것이 아니며 미래의 상호작용을 고려한 투자이자 보험임을 알 수 있다. 사람들은 좋은 평판을 쌓거나 나쁜 평판을 피하기 위해서 이타적으로 행동한다. 왜냐하면 평판은 보상과 처벌의 효과를 모두 가지고 있기 때문이다. 좋은 평판은 미래의 상호작용에서 도움을 받을 확률을 높이고 짝이나 동료로 선호되게 함으로써 보상의 효과를 가지고 있으며, 나쁜 평판은 미래 상호작용에서 소외될 가능성을 높임으로써 처벌의 효과를 가진다. 즉 평판은 지연된 보상이자 처벌의 특성을 가지고 있는 것이다. 중요한 점은 이러한 평판은 보상이나 처벌에 비해 비용이 거의 들지 않으며 구성원 간에 쉽게 공유되기 때문에 기존 인간의 이타적 행동을 설명하는 이론들의 한계를 보완하며 효과적으로 집단 내 이타적 협동을 유지할 수 있는 요소로 작동할 수 있다는 것이다.

위와 같은 이점에도 불구하고 기존의 연구는 평판의 특성을 충분히 고려하지 않고 평판 요소를 실험에 도입하였다. 기존의 연구가 간과한 평판의 특성은 실제 사람들이 내린 평가가 아닌 연구자가 임의로 설정한 평판 점수를 사용하였다는 점, 긍정적 평판의 상승폭과 부정적 평판의 상승폭에 차이가 없다는 점, 자신의 협동 여부에 따라 상대방에게 부여하는 평판 경향이 달라질 수 있다는 것을 고려하지 않았다는 점, 공공재 게임의 성공 여부에 따라 달라질 수 있는 평판 부여 경향의 차이를 고려하지 않았다는 점, 자원 분배의 불평등과 상관없이 모든 사람에게

똑같은 평판 평가 체계를 적용했다는 점에서 찾을 수 있다.

따라서 이 논문에서는 긍정적 평판과 부정적 평판을 분리시켜 도입한 공공재 게임을 통해 실제 사람들이 평판 평가를 하도록 함으로써 상대방을 평가할 때 이타적 행동에 대한 긍정적인 평판보다 이기적 행동에 대한 부정적 평판이 더 강하게 나타남을 증명하였다.

또한 게임에 참여한 상대에 대해 상대방이 게임에서 협동했는지 여부뿐만 아니라 자신의 협동 여부, 게임의 성공 여부, 상대적 자원차이가 평판 평가에 영향을 미칠 수 있는 요인임을 확인하였다. 이 후 자신의 협동 여부에 따라 상대방의 협동 여부에 대한 평판 경향이 다르게 나타남을 증명하기 위하여 자신의 협동 여부와 상대방의 협동 여부에 따라 협동-협동 상황, 협동-배신 상황, 배신-협동 상황, 배신-배신 상황의 총 4가지 상황을 분리하여 상대방에 부여한 긍정적 평판과 부정적 평판을 측정하고 비교하였다. 사람들은 자신과 상대방 모두 협동한 상황에서만 다른 상황에 비해 상대방을 더 긍정적으로 평판하였고, 상대방의 협동 여부에 상관없이 자신이 협동하였을 때보다 자신이 배신하였을 때 더 부정적으로 평판하였다.

다음으로 자신이 참여했던 공공재 게임의 성공 여부에 따라 달라지는 상대방의 협동 여부에 대한 평판 경향을 살펴보았다. 게임의 성공 여부와 상대방의 협동 여부에 따라 성공-협동 상황, 성공-배신 상황, 실패-협동 상황, 실패-배신 상황의 총 4가지 상황을 분리하여 긍정적 평판과 부정적 평판의 부여 경향을 각각 비교하였다. 상대방에 대한 긍정적 평판의 경우 공공재 게임이 성공했을 때 실패한 경우보다 강하게 나타났다. 특히 게임이 성공하고 상대방도 협동하였을 때 긍정적 평판이 가장 강하였으며, 게임이 성공하면 상대방이 배신하더라도 게임이 실패하였을 때보다 상대방에게 더 많은 긍정적 평판을 주었다. 상대방에 대한

부정적 평판의 경우 상대방의 협동 여부와 상관없이 참여한 공공재 게임이 실패하였을 경우에 성공하였을 경우보다 강하게 나타났다.

추가적으로 같은 규칙을 가진 공공재 게임에서 차등적으로 현금을 지급함으로써 자원의 불평등이 심화된 상태에 따른 평판 평가 경향을 탐구하였다. 상대적으로 많은 자원을 가진 상태는 협동하더라도 긍정적 평판을 받기 어렵게 만들고 부정적인 평판을 많이 받게 함으로써 부정적 평판 평가 편향을 강화시켰다. 자원을 적게 가진 상태가 가장 부정적 평판을 받는 경향이 약할 것으로 예상하였으나 실험의 참가자들은 자신보다 자원을 적게 가진 상대보다 자신과 같은 수준의 자원을 가진 상대방에게 더 긍정적으로 평판하고 덜 부정적으로 평판함으로써 자신과 자원 수준이 비슷한 상대방에게 가장 자비로운 평판 경향을 가졌음을 알 수 있었다.

마지막으로 각 요인들이 상대방에 대한 평판 평가 경향에 미치는 영향력을 살펴보았다. 상대방의 자원이 나와 같거나 나보다 적을 경우, 게임이 성공할 경우, 자신이 협동할 경우, 상대방이 협동 할 경우 순으로 상대방에 대한 긍정적 평판을 부여하는 경향이 강하게 나타났고, 상대방의 자원이 나보다 많을 경우, 자신이 배신할 경우, 게임이 실패할 경우, 상대방이 배신할 경우 순으로 상대방에 대한 부정적 평판을 강하게 부여하는 경향이 있음을 알 수 있었다.

주요 용어: 이타적 행동, 공공재 게임, 평판 평가, 자원 불평등

학 번: 2010-22986

<목차>

Ⅰ. 서론

인간 외에도 이타적 행동을 보여주는 동물은 많다. 이 때 각 개체들은 표면적으로 자신의 이익을 포기하고 이타적으로 행동하는 것처럼 보인다. 다른 개체를 속이거나 배신함으로써 자신의 이득이 커질 수 있음에도 불구하고 이타적 행동을 통해 상대의 이익을 상승시키기 때문이다. 따라서 각각의 개체가 가진 생존과 번식이란 목적을 달성하기에 이타적 행동은 다소 모순적으로 보인다. 이 문제를 설명하기 위하여 초기 연구자들은 친족 선택설(Hamilton, 1964)과 상호호혜의 원리(Trivers, 1971)를 적용하였다. 개미나 벌과 같은 사회적 곤충들의 경우 한 군집 내 개체들은 높은 수준으로 유전자를 공유하기 때문에 친족간의 협동을 통해 공통 유전자 보존 가능성을 높임으로써 이익을 취한다. 흡혈 박쥐의 피 나눠주기는 상호 호혜주의를 바탕으로 하는 굶주림에 대비한 보험으로 이해할 수 있다. 인간의 이타적 행동 또한 위의 두 가지 형태의 이타적 행동을 포함한다.

그러나 인간의 이타적 행동은 다른 동물들과 분명한 차이점이 있다. 첫째, 인간의 이타적 행동은 친족의 범주에 국한되지 않는다. 인류의 조상들은 공통 유전자를 많이 가진 친족에게만 이타적 행동을 하지는 않았으며 비친족원들과도 이타적으로 협동하는 전략을 발전시켰다. 둘째, 많은 경우 이타적 행동을 제공하는 대상과 수혜자 사이의 미래에 벌어질 상호 호혜가 불확실하다. 인간은 미래에 아무런 이득이 없을 것을 알면서도 때로는 아무 조건 없이 호의를 베풀기도 한다. 무조건적으로 보이는 이타적 행동은 상이한 문화와 생산 체제 전반에 걸쳐서 나타난다. 수렵 채집 사회에서 나타나는 음식 나누기는 대부분 일방향적으로 이루어진다. 파라과이의 아체 족이나 동아프리카의 핫자 족의 경우 사냥해온 고기는 무조건적으로 공동체의 대부분 혹은 전체 인원들과 공유된다. 음식을 더 많이 생산하고 지속적으로 공유하는 사람은 정해져 있으며 이들은 나눠준 사람들로부터 돌아올 음식을 기대하지도 않는다(Hawkes, 1992,1993; Kaplan &

Hill, 1985). 이러한 이타적 행동은 미리암족의 거북이 사냥에서도 발견할 수 있으며(Bird et al. 2002), 보다 큰 규모의 형태로 말라네시아의 정치적 지도자인 빅맨의 연회, 북미 북서 해안 인디언의 포틀래치, 자본주의에서의 자선활동 등으로 나타난다. 현실을 단순화하여 반영한 실험에서도 인간의 이타적 성향을 발견할 수 있다. 익명으로 단 한번 진행되는 독재자 게임[1]을 할 때도 사람들은 자신의 이익을 어느 정도 포기하고 이익을 나누며 이러한 현상은 다양한 문화권에 걸쳐 나타난다(Roth, et al., 1991).

인간은 왜 손해로 보이는 이타적인 행동을 하는가에 대한 문제를 진화 메커니즘을 기반으로 해결하기 위하여 비싼 신호 이론(Smith & Bird, 2005), 이타적 처벌(Boyd, et al., 2005; Fehr & Fischbacher, 2004), 간접 호혜주의(Alexander, 1986), 평판 형성 이론 (Bereczkei, et al.,2007,2010)등이 기존 연구가 있었다. 본 논문은 이타적 행동의 진화적 이점을 증명하는 앞의 연구들과 맥락을 같이 하며, 평판의 개념이 이타적 행동을 설명하는데 가지는 이점을 설명하고 각 이론들 간 연결고리를 찾는다. 또한 평판 평가를 도입한 공공재 게임을 통해 이타적 행동에 대한 긍정적 평판보다 이기적 행동에 대한 부정적 평판이 더 강하게 나타나는 부정적 평판 평가 편향이 존재함을 제안한다. 나아가 평판 평가에 영향을 미치는 추가적 요소들로 자신의 협동 여부, 게임의 성공 여부, 자원의 불평등을 설정하고 각 요소들이 상대방에 대한 평판 평가 과정에 미치는 영향을 살펴볼 것이다.

[1] 독재자 게임: 참여자 A는 초기 일정한 금액을 받고 참여자 B에게 자신이 가진 일정 부분을 나눠 줄 수 있다. 그러나 B가 A의 제안을 수락하거나 거부할 수 있는 최후 통첩 게임과 달리 B는 A의 제안을 거부할 수 없다.

Ⅱ. 이론적 배경

1. 평판과 타인들

평판은 자신에 관한 것이지만 주위 사람들이 부여하는 것이기 때문에 스스로 통제하기 어렵다. 따라서 이타적 행동을 평판을 통해 이해하기 위해서는 먼저 주위 사람들의 존재가 이타적 행동에 미치는 영향을 살펴볼 필요가 있다. 실제 사람들은 주위 사람들의 존재에 매우 민감하며 비공식적일 때보다 공식적일 때 이타적 행동을 많이 한다(Bereczkei, et al., 2007). 뿐만 아니라 사람들의 이타적 행동을 측정하는 실험에서 참가자들은 다른 사람들이 존재할 것 같은 상황에서 훨씬 관대한 분배 성향을 보이며, 실제 사람의 모습이 아닌 눈 모양의 그림조차 사람들을 이타적으로 행동하게 만드는 효과가 있었다(Bateson et al., 2006; Haley & Fessler, 2005). 최근에는 더욱 단순한 그림을 사용하여 사람들이 얼마나 다른 구성원의 존재에 민감한지 측정하였는데, 점 세 개가 역삼각형 모양으로 찍혀있는 얼굴을 가장 단순하게 표현한 그림까지도 사람들을 더 너그럽게 만드는 것이 밝혀졌다(Rigdon, et al., 2009). 이타적 행동의 질 또한 타인의 유무에 따라 달라진다. 이타적 행동의 비용이 비쌀수록 타인의 존재 유무는 더 중요해진다. 시간, 노력, 자원 등의 비용이 많이 드는 이타적 행동일수록 공식적일 때 더 많이 나타나며 반대로 이타적 행동의 비용이 적게 드는 것은 비공식적일 때 잘 나타난다(Bereczkei, et al., 2010). 또 다른 실험에서는 참가자들이 실험 중에 느낀 생각을 살펴보았는데 지켜보는 눈의 존재는 참가자들을 더 너그럽게 행동하도록 만들었으며, 그들의 너그러운 행동이 나중에 보상 받을 수 있을 것이라는 생각이 들도록 만들었다(Oda et al., 2011). 궁극적으로 인간의 이타적 행동은 관찰자의 존재라는 특정 조건에 따라 다르게 나타나며 이는 인간이 항상 이타적으로 행동하는 무조건적 이타주의자가 아님을 알려준다. 다시 말하자면 인간의 이타적 행동은 의식적이든 무의식적이든 매우 경제적이고 효과적인 보험이자 투자인 것이다.

2. 비싼 신호 이론과 긍정적 평판

비싼 신호이론은 비싸고 낭비로 보이는 행동이나 형태적 특질이 정직한 신호로서 효과적으로 전달되기 위해 디자인되었다는 이론이다(Zahavi, 1975). 사슴의 뿔과 마찬가지로 인간의 이타적 행동 또한 개인의 내재된 자질을 드러낸다. 모든 행동에는 시간, 에너지, 위험, 자원이 따른다. 따라서 이타적 행동의 경우 비용 대 편익을 표면적으로 계산해 볼 때 이해타산이 맞지 않는 낭비이다. 하지만 이처럼 비싸고 위험할 수 있는 행동은 비싸고 낭비적이라는 바로 그 이유 때문에 개인의 자질을 들어낼 수 있는 정직하고 효과적인 수단이며 주변도 그 신호에 민감하게 반응하는 것이다. 수렵 채집 사회의 음식 나누기에서 나타나는 이타적 행동도 비싼 신호이론을 통해 이해할 수 있다. 이타적 행동은 자신의 자질을 과시함으로써 행위자의 지위를 높일 수 있는 효과적 수단이며(Smith & Bird, 2005), 매력적인 짝이나 동지로서 인식되어 미래의 상호작용에서 이익을 얻게 한다(Barclay & Willer, 2007). 수렵 채집 사회의 이타적 음식 나누기뿐만 아니라 산업사회에서 나타나는 낯선 사람에게 자선을 베풀고 기부하는 행동 또한 행위자의 성격적 특성과 협동성을 나타내는 비싼 신호이다(Bereczkei, et al., 2010). 따라서 이타적 행동은 단기적인 비용에도 불구하고 이타주의자의 사회적 지위를 높임으로써 이타주의자가 자원에 접근할 수 있는 가능성을 높이며, 미래의 상호관계에서 유리한 위치를 점할 수 있도록 하여 장기적으론 행위자에게 이익을 가져다 준다(Wedekind & Braithwaite, 2002).

비싼 신호이론을 통해 살펴본 이타적 행동의 이점인 높은 사회적 지위 획득과 마찬가지로 좋은 평판을 가진 사람 또한 짝이나 동맹으로 선호된다. 따라서 좋은 평판을 가진 사람 또한 미래의 상호관계에서 유리한 위치를 점할 수 있으며 발전적 교류를 통해 자원에 접근하기도 쉽다. 이러한 이점들이 사람들에게 좋은 평판을 형성하고 싶은 욕구를 유발시키며 결과적으로 이타적 행동을 촉진하는 것이다.

앞서 살펴보았듯이 비싼 신호 이론과 긍정적인 평판 형성은 이타적 행동의 이점을 설명하는데 있어서 어느 정도 비슷한 측면을 갖는다. 이 때 비싼 신호 이론이 수신자와 발신자의 관계에 초점을 맞춰 수신자의 내재된 자질을 설명하고 신호의 진화적 이점을 설명한다면, 평판은 수신자끼리 공유하는 발신자에 관한 정보이다. 따라서 평판은 발신자의 신호를 확장시킨다. 평판을 통해 직접적으로 발신자의 신호를 관찰한 대상뿐 아니라 다른 구성원들도 간접적으로 발신자의 신호를 공유할 수 있기 때문이다. 평판을 통한 정보교환은 발신자에 대한 정보를 간접적으로 확장함으로써 이타적 행동을 하는 사람과 그 집단 구성원들이 효율적으로 신호를 보내거나, 받아들이고, 공유할 수 있도록 한다.

따라서 좋은 평판을 쌓는 것의 이점은 신호의 직접적인 수신자뿐만 아니라 신호의 간접적 수신자를 통해서도 발생한다. 좋은 평판은 집단 구성원과 상호작용에서 낯선 타인의 도움을 받을 기회를 높이기 때문이다(Wedekind & Milinski, 2000). 때문에 사람들은 다양한 전략으로 집단 내에서 좋은 평판을 쌓기 위해 노력한다. 예로 사람들은 전략적으로 집단의 공공재에 기여하며(Bereczkei, et al., 2007) 나아가 자신이 속하지 않은 집단에서도 자비로운 행동을 한다(Milinski, et al., 2002,2005). 사람들은 이타적 행위에 관한 평판이 어디에서 형성되었든 차별하지 않고 적극 활용하기 때문이다. 내부 집단뿐 아니라 외부 집단에서 이타적 행동을 한 사람도 그가 속한 집단에서 좋은 평판을 얻을 수 있으며 이는 행위자의 정치적 평판을 높이는데 강력한 영향을 미친다(Milinski, et al., 2002,2005). 따라서 어느 한 곳에서 형성된 좋은 평판은 그 집단에 한정되지 않고 다른 집단에서도 이익을 얻을 수 있게 함으로써 확장된 보상의 효과를 가지고 있음을 알 수 있다.

3. 이타적 처벌과 부정적 평판

집단의 구성원들을 기만하여 결과적으로 집단 공공재 생산에 방해가 되는 이기적인 개인은 언제나 존재한다. 집단의 이익과 상충되는 개인의 이익을 취할 수 있는 상황은 어느 곳에나 존재하며 위반으로 인해 얻을 수 있는 이익의 유혹은 강력하기 때문이다. 처벌은 집단 행동을 통해 발생하는 협력자의 이익보다 무임승차자의 이익이 크지 않도록 방지한다. 따라서 규범 위반자에게 가해지는 처벌은 사회가 유지되도록 돕는 필수 요소라 할 수 있다(Fehr & Gächter, 2002). 처벌하는 사람과 규범 위반자의 관계에 따라 제 2자의 처벌과 제 3자의 처벌을 구분할 수 있다. 제 2자 처벌은 상대방이 그 자신에게 해를 가했을 경우 직접 위반자에게 보복하는 것을 뜻하며 제 3자 처벌은 자신이 아닌 다른 사람에게 해를 가한 대상을 처벌하는 것을 뜻한다. 이 때 제 3자는 자신이 직접적인 피해를 입지 않았고, 처벌을 하는데 위험과 비용을 감수해야 하기 때문에 이타적 처벌자, 강한 호혜주의자, 비싼 처벌자 등으로 불린다(Fehr & Gächter, 2002; Boyd, et al., 2005; Fehr & Fischbacher, 2004). 작은 집단의 사회일 경우 집단 내 상호 협력을 증진하는데 제 2자의 처벌로 충분할 수 있다(Marlowe, et al., 2008). 그러나 집단이 커지고 구성이 복잡해 질수록 위반과 은폐가 쉬워지기 때문에 직접적 처벌의 한계가 발생할 수 밖에 없다. 이때 이타적 처벌은 집단이 커질수록 자주 접하게 되는 집단행동의 딜레마를 해결하는데 필수적이라 할 수 있다(Boyd, et al., 2005). 실제로 비교 문화 연구를 통해 집단의 크기가 크고 구성이 복잡해질수록 더 강한 이타적 처벌이 나타남을 알 수 있다(Marlowe, et al., 2008).

그러나 이타적 처벌만으로 대규모의 협동을 안정적으로 유지하기는 쉽지 않다. 왜냐하면 처벌하는데 필요한 비용으로 인해 구성원 전체의 평균 이익은 줄어들 뿐만 아니라, 자신의 비용으로 무임승차자를 처벌하지 않으려는 제2차 무임승차자

문제를 해결해야 하기 때문이다(Panchanathan & Boyd, 2004; Ohtsuki & Iwasa, 2009). 평판은 궁극적으로 협력을 불안정하게 만드는 위의 두 문제를 어느 정도 해결할 수 있다는데 중요성을 가진다. 평판은 대상의 긍정적 자질뿐만 아니라 부정적 자질도 포함할 수 있다. 사람들은 이타적 행동을 하는 사람에게 좋은 평판을, 이기적 행동을 하는 이에게 나쁜 평판을 가한다(Bereczkei, et al., 2007,2010). 이 때 나쁜 평판을 받은 개인은 타인의 도움을 받을 확률이 줄어들거나(Sommerfeld, et al., 2007) 동료로 선택되기 힘들어지는 등 다른 구성원들과 미래의 상호작용에서 불이익을 받게 된다.

이타적 처벌과 평판은 부정행위를 방지하고 직접적인 피해자가 아닌 주위 구성원이 무임승차자에게 손해를 입힌다는 점에서 공통점을 가진다. 실제 사람들이 부정적 평판을 할 때와 이타적 처벌을 할 때 느끼는 정서적 동기는 같다(Feinberg, et al., 2012a). 그러나 평판은 처벌에 비해 비용이 거의 발생하지 않는다. 따라서 이타적 협동을 유지하기 위한 개인들의 가장 효과적인 전략은 처벌을 내리기보다 평판을 통해서 이기적인 개인에게 도움을 주지 않고 상호작용에서 소외시키는 것이다(Panchanathan & Boyd, 2004; Ohtsuki & Iwasa, 2009). 결과적으로 사람들은 미래에 발행할 상호작용에서 소외 당하지 않기 위해 이타적으로 행동하며 이는 평판이 선행을 장려하고 악행을 방지하는 수단이 될 수 있음을 증명한다 (Sommerfeld, et al., 2007). 그러나 이타주의적 처벌이 평판에 의한 차별 전략으로 완전히 대체되지는 않는다. 무임승차자들은 처벌과 평판이 결합된 환경에서 가장 심하게 처벌받기 때문이다(Rockenbach & Milinski, 2006). 결국 처벌과 평판 체계가 조화를 이룰 때 대규모의 협동이 유지될 수 있으며 처벌과 평판 형성 사이의 상호작용은 집단 내 협동의 효율성을 높인다고 할 수 있다.

4. 평판의 작동과 측정

이타적 행동을 계측된 평판을 통해 이해하려는 시도는 간접적 호혜주의의 문제를 해결하기 위해 시작되었다. 간접적 호혜주의란 이타주의자를 도와줌으로써 이타적 행동이 진화해왔다(Alexander, 1986)는 이론이다. 비싼 신호 이론에서는 신호를 보내는 사람이 장기적 이익을 얻기까지 발생하는 모든 비용을 부담하고 수신자가 신호를 이용하더라도 그 대가를 지불할 필요가 없다. 반면 간접 호혜주의에서 행위자가 얻는 이득은 미래의 상호작용에서 만난 타인의 비용으로부터 발생한다. 즉, 간접 호혜주의는 자신이 도왔던 사람이 아닌 타인이 나를 도움으로써 결과적으로 지연된 호혜주의가 발생하는 것이다. 또한 이는 나를 도운 사람을 돕는 것이 아니라 남을 도왔던 사람을 돕는다는 점에서 직접적 호혜주의와도 구별된다. 이러한 아이디어를 발전시켜 보이드Boyd와 리처슨Richerson은 사람들이 원 모양의 순환 형태로 도움을 주고 받으면 간접적 호혜주의가 효과적일 수 있음을 증명하였다. 즉 A가 B를 도우면 C는 A를 돕고 B는 C를 돕는 것이다(Boyd & Richerson, 1989). 그러나 이 방법은 아주 작은 집단에만 적용된다는 한계가 있었다. 직접적 호혜주의의 성공도 어려운 상황에서 간접적 호혜주의는 사용자에 대한 지적 요구 사항도 많을 뿐만 아니라 안정적 사회환경도 필요하기 때문에 유동적이고 규모가 큰 집단에서 제 3자에 의한 보답의 순환이 어렵기 때문이다(Pollock & Dugatkin, 1992). 노바크Nowak와 지그문트Sigmund는 시뮬레이션[2]의 모든 참가자에게 이미지 점수를 주는 방식으로 평판 요소를 도입하여 앞의 문제를 해결하며 간접적 호혜주의가 효과적임을 증명하였다. 모든

[2] 100명을 구성으로 한 컴퓨터 시뮬레이션으로 참가자 100명을 무작위로 둘씩 짝지어 50쌍을 만들고 한 쌍 중 한 명에게 짝을 도울지 선택하도록 한다. 이때 짝을 돕겠다고 하면 참가자의 적합성 점수는 약간 떨어지고 짝의 점수는 크게 올라가며, 돕지 않으면 둘의 점수는 변하지 않는다. 그리고 두 참가자는 다시 만나지 않고 짝을 몇 차례 바꾸어가며 실험하여 모두 선택할 기회가 돌아간 후 번식을 시작하여 다음 세대에는 진화적합성 점수가 높은 개체가 증가하고 낮은 개체는 밀려나게 한다.

사람은 0점에서 시작하며 연구자는 참가자들의 상호작용을 관찰하여 참가자가 도움을 줄 때는 1점을 추가하고, 도움을 주지 않을 때는 1점을 하락시켰다. 이를 통해 참가자들은 다른 참가자의 과거 행동을 알 수 있는데 만약 짝의 이미지 점수가 양수라면 대부분 도왔다는 것이고 음수라면 그 반대라는 것이다. 진화적합성이 높은 개체만 살아남게 하는 프로그램에서 세대가 거듭할수록 이미지 점수가 0점 이상인 참가자만 도와주는 전략이 가장 효과적인 전략으로 드러났으며 이는 평판으로 협력이 증가하며 배신이 소멸할 수 있음을 보여준다(Nowak & Sigmund, 1998). 앞의 연구와 같이 참가자가 짝을 도와주는지 여부에 따라 이미지 점수 1점을 추가하거나 하락시키는 방법을 사용하여 실제 사람들에게 직접 짝을 도울지 말지 결정하는 실험에서 이미지 점수가 양수를 기록한 참가자는 짝에게 돈을 받을 확률이 훨씬 높았다(Wedekind & Milinski, 2000).

실제 사회에서 평판은 소문을 통해 매우 빠르고 효과적으로 퍼진다. 일반 사람들의 대화의 평균 65%는 그 자리에 없는 제 3자에 대한 소문으로 이뤄져 있기 때문이다(Dunbar, 2004). 이러한 소문은 그 신뢰성이 낮더라도 상호작용을 하는 대상에게 매우 민감하게 받아들여진다. 한 실험에서 참가자들에게 긍정적이거나 부정적인 소문을 알려주었을 때 그 소문의 신뢰성이 낮더라도 참가자들은 그 소문에 따라 의사결정의 방향을 바꿨다(Sommerfeld et al., 2007). 따라서 사람들은 자신의 이기적 행동에 대한 소문이 퍼질 수 있다는 것을 알 때 의식적으로든 무의식적으로든 더 이타적으로 행동하게 된다(Sommerfeld et al., 2007; Piazza & Bering, 2008). 즉 사람들은 좋은 평판을 증가시키기 위해서(Hardy & Van Vugt, 2006) 혹은 이타적인 행동을 하지 않았을 경우 얻게 될 나쁜 평판이 두렵기 때문에(Hoffman, et al., 1994) 이타적으로 행동하며 이는 곧 대규모의 협동을 유지하는 기반이 된다. 특히 규모가 작은 사회의 경우 집단 전원이 소문에 쉽게 공유할 수 있기 때문에 소문은 이타적 협동을 유지하는 수단으로서 더욱 효과적이다(Feinberg, et al., 2012b).

Ⅲ. 문제 제기 및 연구 목적

인간의 상호작용에서 평판이 중요한 이유는 한 사람이 과거에 얼마나 베풀었는가를 나타내는 지표일 뿐만 아니라 그 사람이 앞으로의 상호작용에서 이타적으로 행동할 것인지 예측할 수 있게 하기 때문이다. 이러한 평판은 쉽게 공유되고 행위자의 행동에 따라 부정적이거나 긍정적으로 변하는데 비용이 거의 들지 않기 때문에 이타적 행동을 진화적 관점으로 이해하려는 기존 이론의 한계를 보완할 수 있는 장점이 있다.

그러나 기존의 연구들이 평판 요소를 경제적 전략 게임에 도입할 때 실제 평판의 특성을 간과하였다. 기존 이미지 점수를 사용하여 평판을 도입한 연구에서는 모든 참가자들의 이미지 점수를 0점에서 시작하게 하여 연구자는 참가자들끼리의 상호작용을 관찰하며 참가자가 도움을 줄 때 점수 1점을 추가하고, 도움을 주지 않을 때 1점을 하락시켰다(Nowak & Sigmund, 1998; Wedekind & Milinski, 2000). 이러한 기존의 평판 측정 방법은 오랜 기간 쌓아온 좋은 평판이 한 순간에 무너질 수 있는 실제 사회의 평판 체계와 거리가 멀다. 뿐만 아니라 실제 집단 내 구성원 간이나 인터넷상의 특정 대상에 대한 소문이나 평가는 부정적인 경우가 많다는 것을 경험적으로 알 수 있다. 평판은 신과 같은 위치의 사람이 공정하게 주는 것이 아니다. 평판은 개인이 속해 있는 집단의 구성원들에 의해 만들어지는 것이기 때문에 이는 충분히 객관적이고 선형적이지 않을 수 있다. 따라서 평판을 통해 인간의 이타적 행동의 문제를 해결하기 위해서는 우선적으로 인간의 이타적이거나 이기적인 행동에 부과되는 각각의 평판의 종류와 강도를 고려하고 평판이 형성되는 과정에 영향을 미치는 요소들을 추가하여 전략적 게임에서 사용될 평판의 측정 방법을 다듬을 필요성이 있다.

이미지 점수를 통해 이타적 행동의 연관성을 설명하려는 기존의 연구들이 간과한 평판의 특성은 다음과 같다. 첫째는 실제 참가자들이 부과한 평판 점수가 아닌 연구자가 임의로 대상의 행동에 대해 획일적인 평판 점수를 부여했다는 점이다. 둘째는 긍정적 평판과 부정적 평판의 동기에 차이를 두지 않고 같은 선상에서 해석함으로써 긍정적 평판의 상승폭과 부정적 평판의 상승폭을 동일하게 설정한 것이다. 세 번째로 게임에서 평판을 받는 대상의 협동했는지 여부만 고려할 뿐 평판을 부여하는 자신의 협동 여부를 고려하지 않았다는 점이다. 네 번째로 공공재 게임의 성공 여부에 상관없이 상대방의 협동 여부만 고려하였다는 것이다. 마지막으로 평판하는 대상과 자신의 상대적 자원 차이와 상관없이 모든 사람에게 동일한 평판 부여 체계가 적용된다는 것이다.

따라서 이번 연구는 실제 참가자들이 직접 평판을 부여하게 함으로써 기존에 간과하였던 평판 형성의 특성을 밝히는데 목적이 있다. 구체적으로는 평판 평가를 도입한 공공재 게임 실험을 통하여 보상보다 처벌이 선호되는 것처럼 이타적 행동에 대한 긍정적 평판보다 이기적 행동에 대한 부정적 평판이 더 강하게 나타나는 부정적 평판 평가 편향이 존재함을 제안한다. 또한 상대방의 협동 여부만을 평판의 기준으로서 삼지 않고 자신의 협동 여부, 게임의 성공 여부, 상대방과의 상대적 자원 차이 등의 요인들을 추가하여 이 요인들이 상대방에 대한 평판 평가 과정에 미치는 영향을 살펴보는데 목적이 있다. 따라서 각 요인들에 따른 상황을 구분하여 각각의 상황에서 나타나는 평판 경향을 비교한 뒤, 각 요인이 평판 과정에 미치는 영향력을 비교한다. 즉 자신의 협동 여부에 따라 달라지는 상대방의 협동 여부에 대한 평가, 게임의 성공 여부에 따라 달라지는 상대방의 협동 여부에 대한 평가, 자원의 상대적 차이에 따른 상대방의 협동 여부에 대한 평가를 살피고 각 요인의 영향력을 분석한 뒤에 각 상황에서 나타나는 참가자들의 평판 경향과 감정 상태를 추적한다.

Ⅳ. 연구 가설

1. 부정적 평판 평가 편향

나쁜 평판은 좋은 평판보다 쉽게 형성되고 공유되며 집단의 협동을 유지하는데 더 선호될 수 있다. 우선 나쁜 평판이 형성되기 쉬운 이유는 인간이 부정적인 것에 대한 집중 편향을 가지고 있기 때문이다. 이러한 집중편향은 위협이나 두려움과 연관된다(Fox, et al., 2002). 온갖 속임수와 배신이 난무하는 사회적 상호작용 또한 두려운 것이 될 수 있다(O¨hman & Mineka, 2001) 무임승차자들의 비용을 지불하지 않은 채 이익을 얻으려는 행동은 이타적으로 협동하려는 사람들의 이익을 위협하기 때문이다(Axelrod, 1984; Cosmides & Tooby, 1992). 상대방의 속임수와 배신이 가져오는 피해는 상당할 수 있기 때문에 진화 메커니즘은 뛰어난 부정행위 탐지 능력을 인간에게 제공했다. 사람들은 사회적인 규범이 위반되었을 때 특히 민감하며(Cosmides & Tooby, 1992) 부정행위자들의 얼굴을 협력자들의 얼굴보다 더 정확하게 기억한다(Chiappe & Brown, 2004). 심지어 사람들은 무의식적으로 비협력자들 얼굴이 더 무섭고 더 위협적이라고 판정하고 차별한다(Vanneste, et al., 2007). 이러한 연구들은 우리의 부정행위 탐지 능력이 협력을 위협하는 개인들 쪽으로 편향되어 있다는 것을 보여준다. 또한 부정적인 정보나 사건은 긍정적인 정보나 사건보다 충격이 더 크고 더 오래 기억된다(Ito, et al., 1998; Peetersa & Czapinskib, 1990). 따라서 평판은 선행보다 부정행위에 민감한 형태로 나타날 가능성이 크다.

부정적인 평판은 공유되기도 쉽다. 사람들은 긍정적인 정보보다 부정적인 정보 혹은 위협이 되는 정보를 더 쉽게 공유한다(Taylor, 1991). 긍정적인 정보의 공유는 자신을 이익이 감소시킬 가능성이 있으나 부정적인 정보를 흘려준다고 해서

자신의 이익이 감소할 가능성은 낮기 때문이다. 타인의 이타성에 관한 정보 또한 같은 맥락에서 해석할 수 있다. 무임 승차자에 대한 정보를 공유한다고 해서 자신의 이익이 줄어들 가능성은 낮으나 이타주의자는 매력적인 파트너이자 짝이기 때문에 사람들은 이타주의자로부터 얻는 이익을 독차지하고자 노력할 수 있다.

공공의 이익을 위해 협동하는 사람이 무임승차자를 처벌하려는 동기는 협동하지 않는 사람이 협동하는 사람에게 보상하려는 동기보다 더 강하다(Price, et al., 2002). 또한 반복되는 공공재 게임에서 참가자들은 처벌을 보상보다 점차 선호할 뿐 아니라 처벌이 보상보다 집단의 협동을 유지하는데 효과적임이 밝혀졌다(Gürerk, et al., 2004). 부정적인 평판을 퍼뜨리게 하는 감정과 처벌을 할 때 느끼는 감정이 같다는 측면에서(Feinberg, et al., 2012a) 부정적 평판을 퍼뜨리는 것이 긍정적 평판을 퍼뜨리는 것 보다 더 선호될 수 있다. 따라서 총 평판이 긍정적 평판에서 부정적 평판을 뺀 값일 때, 이타적 행동에 대한 긍정적 평판의 상승폭 P와 이기적 행동에 대한 부정적 평판의 상승폭 N을 도출하여 비교할 필요성이 있다. 따라서 따라서 가설 1의 내용은 다음과 같다.

가설 1: 사람들은 부정적 평판 평가 편향을 가지고 있다. 즉 이타적 행동에 대한 긍정적 평판의 상승폭 P는 이기적 행동에 대한 부정적 평판 상승폭 N보다 작다.

$R = R_p - R_n \ (P \leq N)$

$R_p = P \cdot r_p + e$

$R_n = N \cdot r_n + e$

R : 총 평판

R_p : 긍정적 평판

R_n : 부정적 평판

r_p : 이타적 행동 (1회)

r_n : 이기적 행동 (1회)

P : 긍정적 평판 상승폭

N : 부정적 평판 상승폭

2. 자신의 협동 여부에 따른 평판 경향

처벌과 보상 모두 대규모의 협동을 유지할 수 있는 하나의 방법이나 사람들은 자신의 협동 여부에 따라 다른 방식으로 협동을 유지하려 한다. 공공의 이익을 위한 협동에 참여 의지가 낮은 사람의 경우 참여하지 않는 사람에게 처벌을 하기보다 참여자에게 보상을 하려는 동기가 강하며, 참여의지가 높은 사람들의 경우 참여자에게 보상을 하는 것보다 참여하지 않는 사람에게 처벌을 가하려는 동기가 더 강하다(Price, et al., 2002). 앞의 이론적 배경에서 살펴보았듯이 평판은 보상과 처벌의 효과를 지니고 있다. 그러므로 공공재 게임에 협동한 개인과 협동하지 않은 개인이 부여하는 평판 또한 차이가 있을 수 있다. 즉 자신의 협동 여부에 따라 상대방의 협동 여부에 대한 평판을 다르게 부여한다고 예상하였을 때, 기존 연구를 바탕으로 무임승차자는 협력자에 비해 다른 무임승차자에게 부정적 평판을 가하지 않고 협력자에게 긍정적 평판을 할 가능성이 크다. 반면 협력자는 다른 협력자에게 긍정적 평판을 주기보다 무임승차자에게 부정적 평판을 가할 가능성이 더 클 것을 예상할 수 있다. 따라서 참가자가 협동했을 때와 배신했을 때를 구분하여 평판 대상의 협동 여부에 따른 긍정적 평판과 부정적 평판이 어떻게 나타나는지 그 차이를 살펴볼 필요성이 있다. 즉 자신이 협동하였을 때 상대방도 협동하는 상황, 자신이 협동하였을 때 상대방은 배신한 상황, 자신이 배신하였을 때 상대방이 협동한 상황, 자신이 배신하였을 때 상대방도 배신한 상황 총 4가지의 상황을 구분하여 부정적 평판과 긍정적 평판이 각각의 상황에서 어떻게 다르게 나타나는지 살펴볼 필요가 있다. 따라서 가설 2의 내용은 다음과 같다.

가설 2: 자신의 협동 여부에 따라 타인의 협동 여부에 대한 평판 경향은 다르게 나타난다. 즉 모든 RP_i는 서로 같지 않으며, 모든 RN_i 는 서로 같지 않다.

$GR_i = RP_i - RN_i$ (i = 1,2,3,4)

GR_i = 부여하는 평판 점수

RP_i = 부여하는 긍정적 평판

RN_i = 부여하는 부정적 평판

i= (자신의 협동 여부, 타인의 협동 여부)

협동: 1, 배신: 0 일 때

1= (1,1)

2= (1,0)

3= (0,1)

4= (0,0)

3. 게임의 성공 여부에 따른 평판 경향

과거의 경제적 전략 게임에서는 주로 인간의 비이성적인 결정에 집중하였으나 인간의 이성적인 계산 또한 무시할 수 없다. 독재자 게임에서 한 푼도 상대방에게 돈을 주지 않는 참가자도 존재하며, 최후 통첩 게임[3]에서 아주 작은 돈이라도 거절할 수 있음에도 불구하고 받아들이는 참가자도 존재한다. 제 3자 처벌 게임[4]에서의 제 3자의 역할을 맡은 사람 중 자신의 돈을 아끼기 위해 규범 위반자에게 처벌을 하지 않는 참가자도 있다. 어떤 참가자들은 경제적 상호작용의 과정만 아니라 게임의 결과를 최우선으로 생각할 수 있으며 다른 행위자들의 상대적 이익에 큰 상관 없이 자신의 이득을 최우선으로 생각할 수도 있다. 결국 인간의 판단은 상호작용의 과정이나 결과 둘 중 하나에만 귀속된 결정이 아니며 복합적인 것일 수 있다. 예를 들어 <실험1>의 게임에서 성공하였을 경우 도토리 1개 혹은 2개의 이익을 얻게 되는데, 자신이 협동한 경우 배신한 참가자가 자신보다 이익이 많다고 해서 게임에 실패하여 도토리를 하나도 못 얻었거나 잃은 경우보다 더 부정적으로 생각하지 않을 수 있다. 또한 게임이 실패한 경우 다른 참가자가 협동하였을지라도 자신의 이익이 발생하지 않기 때문에 상대방을 긍정적으로 생각하기보다 부정적으로 생각할 수 있다. 따라서 게임이 성공했을 때와 실패했을 때를 구분하여 평판 대상의 협동 여부에 따라 상대방에게 긍정적 평판과 부정적 평판을 어떻게 부여하는지 그 차이를 살펴볼 필요성이 있다. 즉 게임이 성공했을 때 상대방도 협동하는 상황, 게임이 성공했을 때 상대방은 배신한 상황, 게임이

[3] 최후 통첩 게임: 참여자 A는 초기 일정한 금액을 받고서 참여자 B에게 자신이 가진 일정 부분을 나눠 줄 수 있다. 이 때 B는 A의 제안을 수락하거나 거부할 수 있다. B가 제안을 수락할 경우 A와 B는 A의 제안대로 금액을 나눠 갖게 되며, 거부할 경우 두 참여자에게 돌아가는 것은 없다.

[4] 제 3자 처벌 게임: 게임의 참여자는 절대자 게임 혹은 죄수의 딜레마 게임을 하는 참여자 A, 참여자 B와 이 둘의 게임 과정을 지켜본 참여자 C로 구성된다. A와 B가 절대자 게임을 진행할 경우 절대자 A는 초기 일정한 금액을 받고서 수혜자 B에게 자신이 가진 일정 부분을 나눠 줄 수 있다. B는 A의 제안을 거부할 수 없으며 이 둘의 게임 결과는 A에게 처벌이 가능한 C에게 제공된다. C는 A보다 적은 일정 금액을 받고 그 돈으로 A에게 처벌을 내려 금전적 손실을 줄 수 있다. 이 때 C 또한 처벌을 내리는데 비용을 지불해야 한다.

실패했을 때 상대방이 협동한 상황, 게임이 실패했을 때 상대방도 배신한 상황 총 4가지의 상황을 구분하여 각각의 상황에서 실험의 참가자들이 부정적 평판과 긍정적 평판을 어떻게 다르게 부여하는지 살펴볼 필요가 있다. 따라서 가설 3의 내용은 다음과 같다.

가설 3: 게임의 성공 여부에 따라 타인의 협동 여부에 대한 평판 경향은 다르게 나타난다. 즉 모든 RP_i는 서로 같지 않으며, 모든 RN_i 는 서로 같지 않다.

$GR_i = RP_i - RN_i$ (i = 1,2,3,4)

GR_i = 부여하는 평판 점수

RP_i = 부여하는 긍정적 평판

RN_i = 부여하는 부정적 평판

i= (게임의 성공 여부, 타인의 협동 여부)

성공: 1, 실패: 0 일 때

협동: 1, 배신: 0 일 때

1= (1,1)

2= (1,0)

3= (0,1)

4= (0,0)

4. 자원의 불평등에 따른 평판 경향

선천적인 것이든 후천적인 것이든 사람마다 가지고 있는 자원은 평등하지 않다. 그리고 각 개인이 가지고 있는 자원에 따라 평판이 부여되는 것에도 차이가 있다. 현실에서 정치인이나 유명 방송인들에게 보통 사람보다 높은 도덕적 기준을 요구하는 상황을 종종 볼 수 있다. 우리는 높은 지위에 있는 사람에게 좋은 평판은 박하며 나쁜 평판은 더욱 가혹하다는 것을 경험적으로 알 수 있다. 지위가 더 높은 사람에게 처벌을 가하는 행동은 지위가 같거나 낮은 사람에게 처벌하는 행동보다 상대적으로 더 큰 비용이 든다. 그러나 평판은 처벌에 비해 많은 비용을 필요로 하지 않는다. 따라서 지위가 낮은 사람이 처벌할 수 있는 가장 효과적인 방법은 집단 다수의 험담을 통한 나쁜 평판을 이용해 간접적인 불이익을 주는 것이다. 이는 궁극적으로 부정적 평판을 하는 것이 지위가 높은 사람이 휘두를 수 있는 잠재적 위험에 대한 대비 수단이 될 수 있음을 의미한다.

좋은 평판 또한 자원의 불평등에 따라 차이가 있을 수 있다. 이타적 행동은 높은 지위에서 많이 나타나는데 높은 지위를 가진 개인의 이타적 행동은 낭비적으로 보이지만 실제 높은 지위를 강화시키는 장기적 이점을 가진다(Boone, 1998). 이타적 행동으로 얻는 이득은 체감 비용 대 편익을 따져봤을 때도 높은 지위를 가진 개인이 유리하다. 많이 가진 자가 선행을 하는데 드는 비용은 적게 가진 사람이 선행을 하기 위해 드는 비용보다 각각이 가지고 있는 자원에서의 비율로 볼 때 적다. 따라서 자원대비 상대적 선행 비용이 적은 사람은 상대적으로 쉽게 선행을 할 수 있다. 신호의 비용이 감소하는 것은 그만큼 신호의 신뢰성이 하락하는 것이다(Smith & Bird 2005; Zahavi & Zahavi, 1997). 이는 많이 가진 자와 적게 가진 자의 선행 비용 체감 정도가 타인들이 그 선행을 평가하는 요소가 될 수 있음을 말한다. 실제로 참가자들이 너그러움이 과시 신호로 사용된다는 사실을 알게 되면 신호의 신뢰성은 하락된다. 즉, 선행의 가치를 평가절하한다(Barcaly & willer, 2007).

결과적으로 비대칭적 자원 분포는 자원을 많이 가진 사람의 이기적 행위에 대한 부정적 평판 상승폭을 증가시키고 이타적 행위에 대한 긍정적 평판 상승폭을 감소시킴으로써 부정적 평판 평가 편향을 강화시킨다. 따라서 자신의 자원과 상대방의 자원이 같은 상황, 자신의 자원이 상대방의 자원보다 적은 상황, 자신의 자원이 상대방의 자원보다 많은 상황으로 총 3 가지 상황을 설정하여 상대방의 협동 여부에 따라 나타나는 평판 경향을 비교할 필요성이 있다. 위의 조건으로 도출한 연구 가설은 다음과 같다.

가설4: 구성원 간 자원의 불평등은 평판 형성 체계에 영향을 미치며 자신과 상대방의 자원 불평등에 따라 각기 다른 긍정적 평판 RP_{id}과 부정적 평판 RN_{id}을 갖는다. 즉 RP_{id}는 모두 서로 같지 않으며 모든 RN_{id} 는 서로 같지 않다. 또한 자원을 많이 가진 상태는 긍정적 평판을 감소시키고 부정적 평판을 많이 받도록 함으로써 부정적 편향을 강화시키는 역할을 한다.

$$GR_{id} = RP_{id} - RN_{id}$$

$i = 0,1$

상대방 협동: 1, 상대방 배신: 0 일 때

GR_{id} = 부여하는 이미지 점수

RP_{id} = 부여하는 긍정적 평판

RN_{id} = 부여하는 부정적 평판

$d = 0,1,2$

0: 자신의 자원 = 상대방의 자원 일 때

1: 자신의 자원 < 상대방의 자원 일 때

2: 자신의 자원 > 상대방의 자원 일 때

Ⅴ. 연구 방법

1. 실험 대상 및 장소

서울대학교에 재학 중인 학생을 대상으로 교내 홈페이지를 통해 지원자 60명(남성 44명 여성 16명, 평균 연령 23.15세 SD 3.384)을 선착순으로 모집하였다. 모집 과정에서 게임을 통해 벌 수 있는 돈 외에 추가적인 참가비가 없음을 공지하였다. 다양한 전공을 가진 피실험자들은 전원 경제적 전략 게임을 해본 경험이 없다.

실험 장소는 사회과학 대학의 전산실이며 참가자들은 다른 참가자들의 존재는 알 수 있지만 참가자 간 거리를 충분하게 두고 칸막이를 설치하였기 때문에 게임을 하면서 서로의 행동을 관찰할 수 없다. 따라서 참가자들은 5명씩 한 팀을 이루어 게임을 하지만 누가 자신과 같은 팀에 속하고 누가 어떠한 의사 결정을 내리는지 서로 알 수 없다.

2. 진행 절차

실험은 2012년 11월 7일과 9일에 총 3번 20명씩 3조로 나눠 진행하였다. 실험은 설명을 포함하여 1시간 10분 정도 소요되며. 게임 방법 설명, <연습게임>, 도토리를 사용하는 <실험1>, 실제 돈을 사용하는 <실험2>로 구성된다. 본격적인 실험을 진행하기 전에 학생들이 게임의 규칙과 내용을 충분히 숙지할 수 있도록 연구자가 준비한 화면을 보며 설명을 듣고 <실험1>과 똑같은 방식의 <연습게임>을 시간 제한 없이 2번, <실험1>과 똑같은 방식으로 시간 제한이 있는 상태에서 1번 진행하였다. 실험 종료 후 자신의 최종적으로 획득한 현금의 액수를 바로 지급하였다.

3. 평판 평가와 공공재 게임

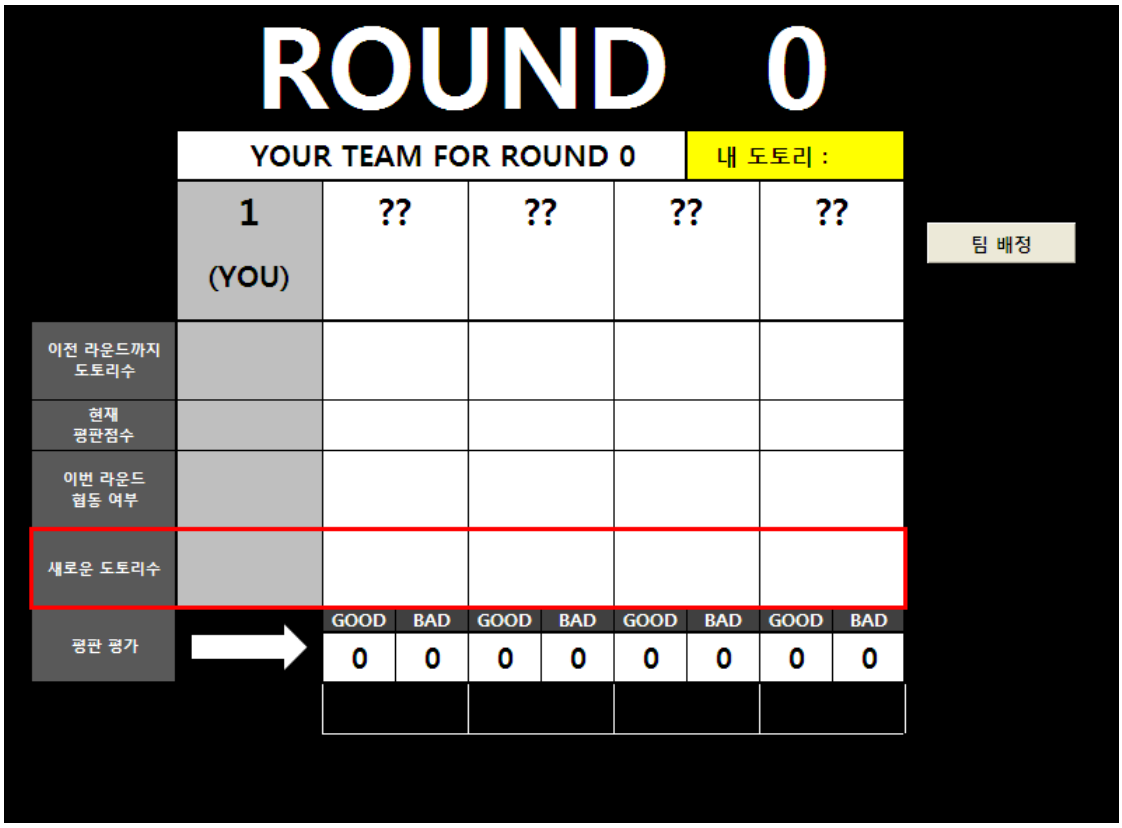
<실험1>

1) 연습게임이 종료 된 후 참가자들은 본격적으로 반복되는 공공재 게임과 평판 평가에 참여한다. <실험1>이 진행되면 각 참여자들이 기존에 가지고 있던 가명은 유지되지만 연습게임의 결과는 모두 삭제되고 초기의 평판과 도토리 상태로 되돌아간다.

2) 학생들은 각각 고유의 가명을 부여 받으며, 중립적인 이미지의 가명을 부여하여 가명이 게임의 진행에 주는 영향을 최소화 시킨다. 가명은 1에서 20까지 아라비아 숫자로 주어진다. 각 참가자가 사용하는 컴퓨터 화면에는 진행되는 공공재 게임의 라운드 수와 공공재 게임에 함께 참여하는 팀의 구성원들의 가명, 공공재 게임을 하기 전의 도토리 수, 직전 라운드에서의 협동 여부, 공공재 게임 후의 도토리 수, 평판 점수가 떠 있도록 한다.

3) 팀 배정을 누르면 참가자들은 무작위로 선출된 5명씩 한 팀이 되어 공공재 게임에 참여한다. 참가자들은 공공재 게임이 한번 끝날 때마다 함께 게임을 한 상대방들에 대한 평판 평가를 할 수 있다. 공공재 게임에서 협동하거나 배신하는 버튼을 누르는 시간은 10초이며 평판 점수를 부여하는 시간은 40초 이내로 벨 소리를 이용하여 제한하였다. 도토리를 사용한 공공재 게임은 총 7번 반복되며 참가자들은 반복되는 공공재 게임이 언제 끝날지 모른다.

그림 1 연습 게임의 초기 화면



4) 각 참가자들은 10개의 도토리를 동등하게 받는다. 참가자들은 협동할 수도 있고 배신할 수 도 있으며 협동에 참여하기 위해서는 개인 당 도토리 1개가 필요하다. 3명 이상 협동할 경우 5명에게 동등하게 2개의 도토리가 지급된다. 3명 미만의 인원만 협동할 경우 모인 도토리는 사라진다. 참가자들은 게임을 하면서 얻은 도토리 개수에 따라 5분위로 나눠 실제 돈을 <실험2>의 게임비용으로 지급받게 된다는 것을 설명을 통해 이미 숙지하고 있다.

그림 2 연습게임에서의 협동과 배신 의사결정 과정

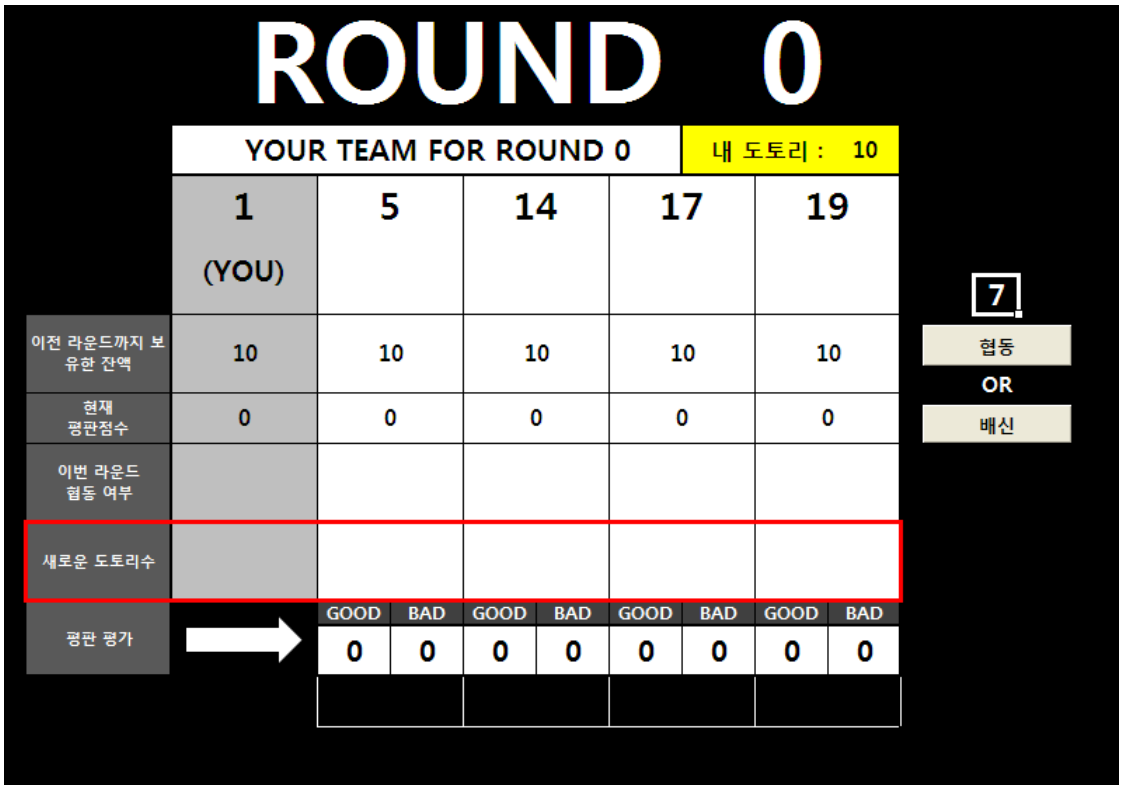
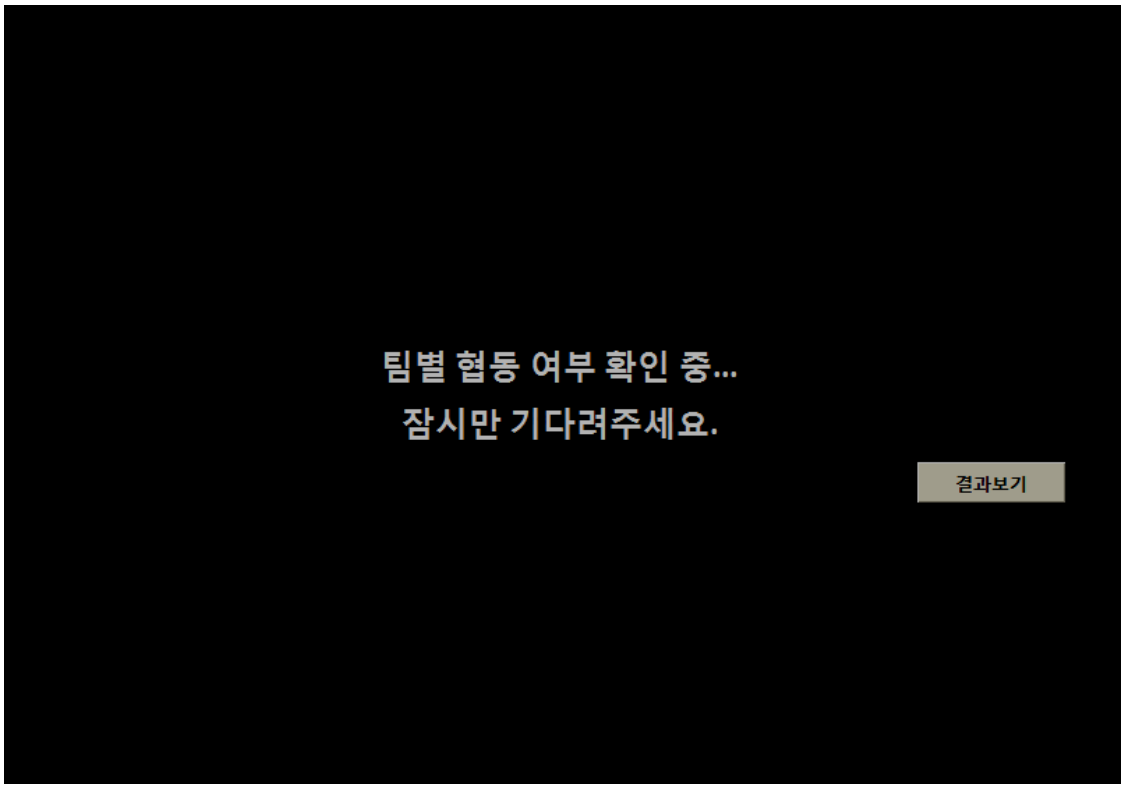


그림 3 팀별 협동 여부 정보 처리 과정



5) 공공재 게임이 한 번 끝날 때마다 참가자들은 서로에게 평판 점수를 부여할 수 있다. 평판은 보상과 처벌의 효과를 가지고 있기 때문에 처벌과 보상의 동기가 다른 것처럼(Price, et al., 2002) 평판 또한 긍정적 평판을 하는 동기와 부정적 평판을 하는 동기에 차이가 있을 수 있다. 따라서 본 연구에서는 긍정적 평판과 부정적 평판을 각각 나누어 참가자들이 직접 상대방에게 긍정적 평판 점수나 부정적 평판 점수를 선택하여 줄 수 있도록 하였다. 이 때 한 명의 상대방에게 긍정적 평판과 부정적인 평판을 동시에 부여할 수는 없다. 한 라운드에서 참가자가 받는 평판 점수는 같은 팀 참가자들이 부여한 긍정적 평판 점수의 총 합에서 부정적 평판 점수의 총합을 뺀 값이다. 만약 한 사람이 얻은 총 긍정적 평판 점수가 2점이고 총 부정적 평판 점수가 6점인 경우 이 사람의 평판 점수는 −4점이 된다. 각 참가자들의 평판 점수는 라운드가 진행됨에 따라 각 라운드에서 받은 평판 점수가 누적된 점수이다.

6) 사람들은 상호작용하는 모든 대상에 대해 똑같은 강도로 평판을 퍼뜨리지 않는다. 평판이 보상이나 처벌보다 비용이 적을지라도 평판을 퍼트리는 것 또한 시간과 노력 등의 작은 비용이 발생한다. 따라서 평판을 내리는 대상이 한정되어 있고 그 강도 또한 한계가 있다. 평판의 대상과 평판 강도를 제한하기 위하여 한 번의 공공재 게임에서 한 명의 참가자가 평판 항목에 기입할 수 있는 점수를 총 10점 이하로 제한한다. 즉 참가자는 총 1점 이상 10점 이하의 점수를 줄 수 있으며 점수를 주지 않을 수도 있다. 만일 한 참가자가 공공재 게임의 결과를 보고 평가할 대상 2명 A와 B를 골라 A의 부정적 평판 항목에 3점을 B의 긍정적 평판 항목에 2점을 부여하였다면 A의 평판 점수는 3점 하락하며 B의 평판 점수는 2점 상승한다.

그림 4 연습게임에서의 평판 점수 부여 과정

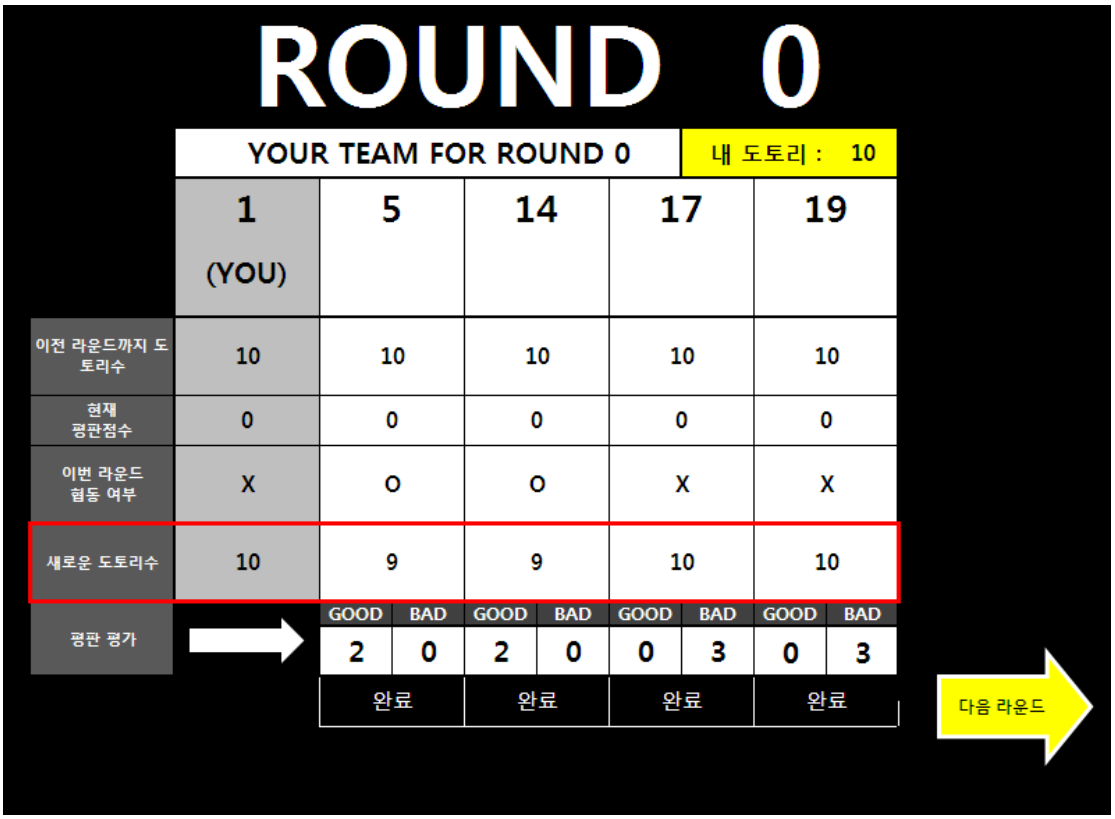
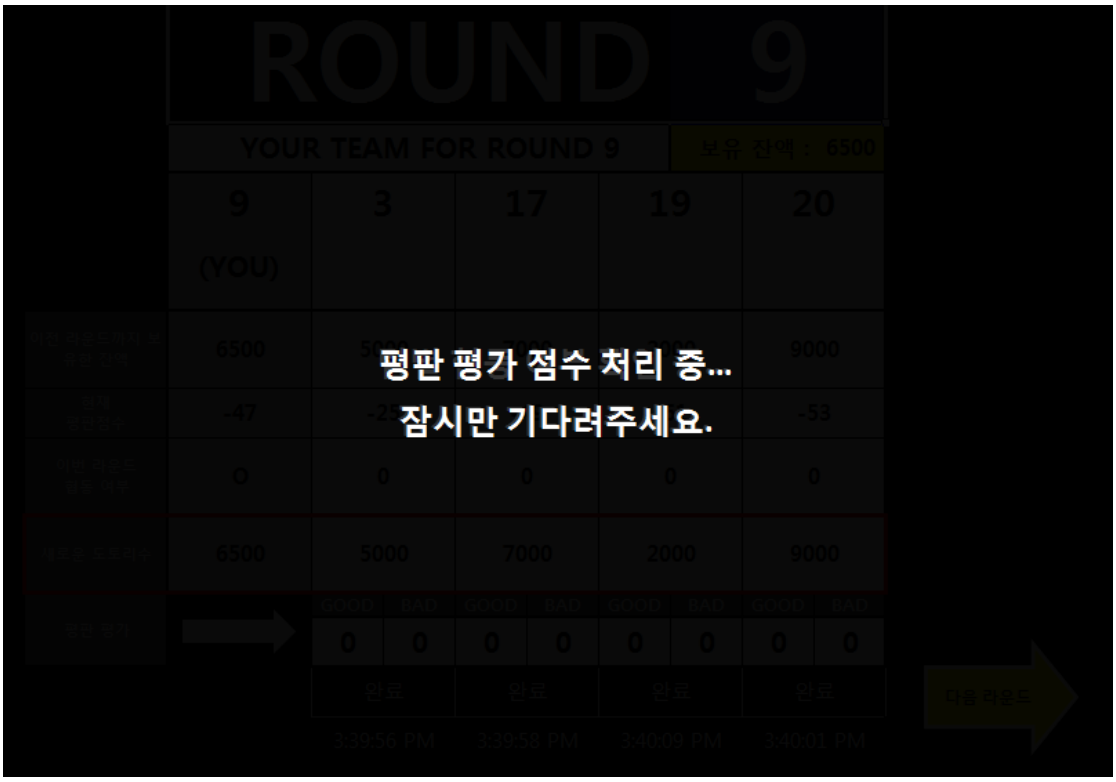


그림 5 평판 점수 정보 처리 과정



7) 좋은 평판의 이점을 실험에 포함시키기 위해 한번 공공재 게임과 평판 평가가 끝날 때마다 가장 비슷한 평판 점수를 받은 사람 5명을 다음 라운드의 공공재 게임 참여자로 구성한다. 과거 실험들에서는 강제적으로 정해진 파트너와 일정기간 게임을 진행하지만 실제 개인은 끊임 없이 협력할 파트너를 찾을 수 있다(Fehr & Fischbacher, 2003). 모든 사람들은 평판이 좋은 사람과 상호작용하기를 원한다. 이 때 평판이 나쁜 사람은 평판이 좋은 사람이 선택해줄 가능성이 낮다. 때문에 평판이 좋은 사람은 평판이 좋은 사람끼리 상호작용할 가능성이 높으며 결국 평판이 좋은 사람 순서대로 한 팀이 될 것이다. 따라서 비슷한 평판 점수를 얻은 참가자끼리 공공재 게임의 새로운 구성원이 되도록 함으로써 현실세계의 평판의 이점을 게임 상황에 도입한다. 즉 평판 점수 입력 후 한 라운드가 끝날 때마다 평판 점수를 5분위로 나눠 같은 분위에 속하는 참가자들끼리 새로운 한 팀이 되도록 한다.

<실험2>

1) <실험1> 종료 후 참가자들에게 자신들의 직전 실험의 결과와 그 결과에 따라 5분위 중 해당되는 분위를 개인에게 공개한 뒤 <실험2>를 시작한다. <실험2>에서는 도토리 대신 실제 돈을 걸고 참가자들은 게임에 참여하며 그 외 실험 내용은 <실험1>과 같다. 참가자들은 <실험1>의 결과에 따라 비대칭적으로 돈을 지급받으며 <실험2> 종료 후 바로 참가자가 최종적으로 획득한 금액을 실제 돈으로 바꿔 받아가도록 한다.

2) 자원의 비대칭성은 참여자들이 받은 가상의 돈의 비대칭성으로 나타난다(김상인,2006). <실험1>이 진행됨에 따라 참가자 간 가진 도토리 수에 차이가 생기긴 하지만 자원의 상대적 차이를 더 심화시키기 위해 <실험2>에서 참가자들은 <실험1>의 결과에 따라 5분위로 나눠 불균등하게 자원을 제공받는다. 즉 <실험1>에서 얻은 도토리에 따라 5분위로 나눠 많은 도토리를 얻은 사람이 많은 돈을 지급받을 수 있도록 한다. 참가자들이 속한 현재 환경과 가장 유사한 자원분포를 설정하기 위하여 2012년 2/4분기 가계동향(통계청, 2012)의 소득 5분위별 소득 금액을 참고한다. 가구를 소득의 금액크기 순으로 배열하여 가구 수를 5등분했을 때 금액이 하위 20%가구가 1분위, 상위 20%가구가 5분위이다. 소득 5분위별 소득 금액은 Ⅰ분위 1275.9, Ⅱ분위 2631.0, Ⅲ분위 3594.4, Ⅳ분위 4716.0, Ⅴ분위 7491.2(단위: 천원)로 나타났다. 이와 비슷한 비율이 되도록 가상의 돈을 Ⅰ분위 1500원, Ⅱ분위 3000원, Ⅲ분위 4500원, Ⅳ분위 6000원, Ⅴ분위 7500원으로 참가자들에게 지급한다.

3) 참가자들은 협동할 수도 있고 배신할 수도 있으며 협동에 참여하기 위해서는 개인당 500원이 필요하다. 3명 이상 협동할 경우 5명에 동등하게 1000원이 지급되며 3명 미만의 인원만 협동할 경우 모인 돈은 증가되지 않고 사라진다.

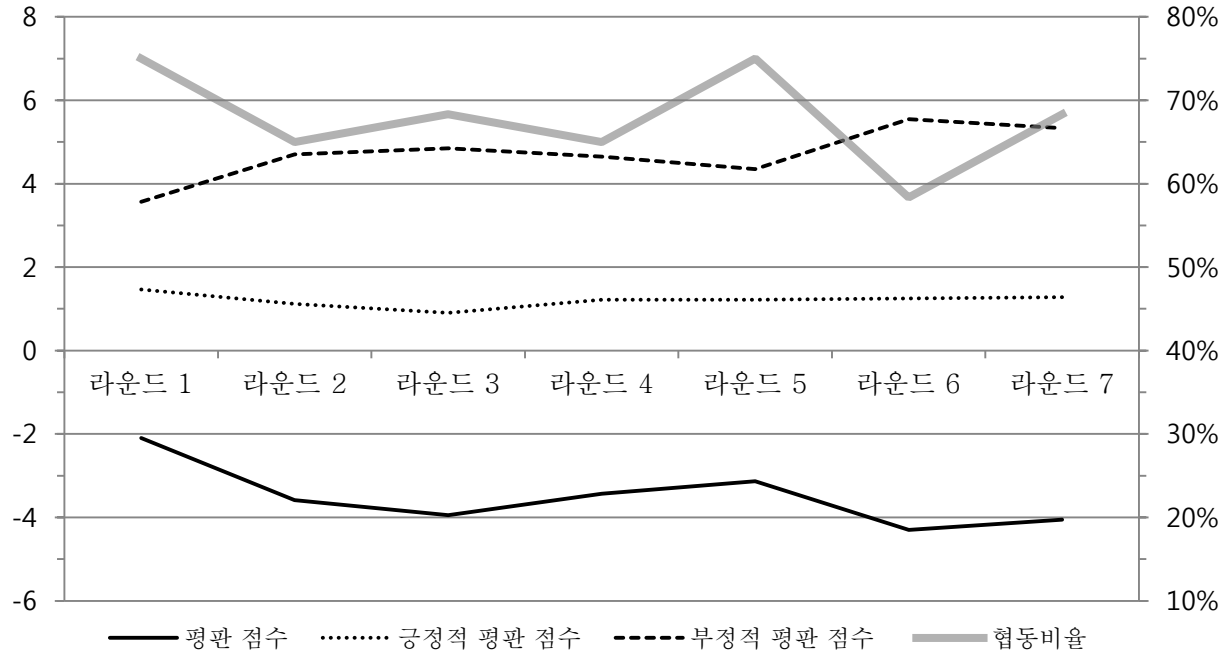
Ⅵ. 결과

같은 수의 도토리로 시작한 <실험1>의 단순통계량은 다음 <표 1>과 같다. 참가자들의 평균 협동 비율은 67.857%이었고, 개인당 획득한 최종 도토리 수는 평균 15.217개이었으며 가장 많이 획득한 사람은 19개, 가장 적게 획득한 사람은 11개로 <실험1>이 종료되었다. 참가자들의 평판 점수는 참가자가 받은 긍정적 평판 점수의 총합에서 부정적 평판 점수의 총합을 뺀 값이며, 참가자들이 받은 최종 평판 점수는 평균 −24.55점이었고 가장 평판 점수가 높은 사람은 25점, 가장 낮은 사람은 −108점이었다. 또한 한 참가자 당 다른 참가자들에게 부여한 총 긍정적 평판 점수의 평균은 16.033점, 총 부정적 평판 평균 점수의 평균은 40.8점이었다. <실험1>의 각 라운드 별 협동 비율과 참가자들이 받은 긍정적 평판 점수, 부정적 평판 점수, 평판 점수의 평균은 <그림 6>과 같다.

표 1 <실험1>의 참가자들의 협동 비율, 획득한 도토리, 최종 평판, 부여한 평판 점수

변수	평균	표준편차	최소값	최대값
회당 평균 협동 비율	0.678	0.309	0	1
참가자가 획득한 최종 도토리 수	15.217	1.86	11	19
참가자가 받은 최종 평판 점수	−24.55	31.23	−108	25
참가자가 부여한 총 긍정적 평판 점수	16.033	16.6519	0	55
참가자가 부여한 총 부정적 평판 점수	40.8	21.1659	4	70

그림 6 <실험1>의 라운드 별 참가자들이 받은 평판의 평균 점수 및 협동 비율

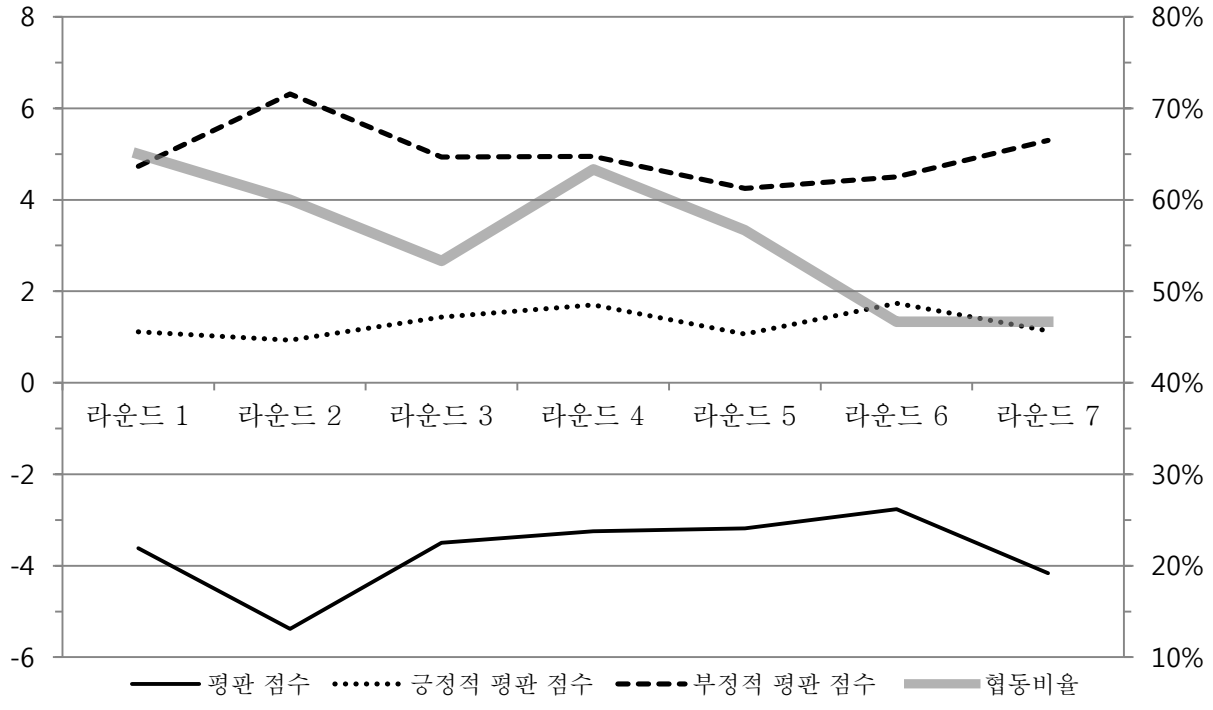


<실험 1>의 결과에 따라 5 분위로 차등적으로 현금을 지급한 <실험 2>의 단순통계량은 다음 <표 2>와 같다. 참가자들의 평균 협동 비율은 55.952%로 <실험 1>보다 협동 비율이 낮은 경향이 나타났다. 게임 종료 후 개인당 획득한 현금은 평균 6792 원으로 현금을 가장 많이 획득한 사람은 11500 원, 가장 적게 획득한 사람은 500 원을 획득하였다. 참가자들은 <실험 1>에 비하여 <실험 2>에서 다른 참가자들에게 부정적 평판을 더 많이 함으로써 공격적인 평판 부여 경향을 보였다. 따라서 참가자들의 최종 평판 점수는 평균 −50.55 점으로 <실험 1>의 최종 평판 점수의 평균보다 낮았으며, 평판 점수가 가장 높은 사람은 55 점, 가장 낮은 사람은 −208 점이었다. 또한 한 참가자 당 다른 참가자들에게 부여한 총 긍정적 평판 점수의 평균은 17.433 점, 총 부정적 평판 평균 점수의 평균은 44.3 점이었다. <실험 2>의 라운드 별 협동 비율과 참가자들이 각 라운드에서 받은 긍정적 평판 점수, 부정적 평판 점수, 평판 점수의 평균은 <그림 7>과 같다.

표 2 <실험2>의 참가자들의 협동 비율, 획득한 현금, 최종 평판, 부여한 평판 점수

변수	평균	표준편차	최소값	최대값
회당 평균 협동 비율	0.559	0.37502	0	1
참가자 1인당 획득한 최종 현금	6792	3129	500	11500
참가자들이 받은 최종 평판 점수	−50.55	50.4819	−208	55
참가자들이 부여한 총 긍정적 평판 점수	17.433	19.1757	0	69
참가자들이 부여한 총 부정적 평판 점수	44.3	21.799	0	70

그림 7 <실험2>의 라운드 별 참가자가 받은 평판의 평균 점수 및 협동 비율



1. 협동 비율과 평판, 수익 및 기타 변인들은 관련성이 있는가?

<실험1>과 <실험2> 에서 협동 비율과 평판 및 수익 사이에 서로 관련이 있는지 살펴보기 위해 SAS 프로그램 ver 9.1.3을 사용하여 유의수준 5% 에서 상관분석(correlation analysis)을 하였다(<표 3>). 피실험자가 많이 협동할수록 긍정적 평판을 많이 받고 부정적 평판은 적게 받음으로써 최종 평판이 더 좋았고 수익도 높았다. 또한 많이 협동하고 평판이 좋을수록 상대방을 더 긍정적으로 평가하며, 협동비율이 낮고 평판이 부정적일수록 부정적 평가를 더 많이 하였음을 알 수 있었다. 수익의 경우 긍정적으로 평판 받고 긍정적으로 평판할수록 높았으며, 부정적으로 평판 받거나 부정적으로 평판할수록 낮은 경향이 나타났다. 성별의 경우 <실험1>과 <실험2>에서 참가자가 남성일 경우 협동비율이 낮으며, 평판 평가에서 부정적 평판을 많이 하고 긍정적 평판을 적게 하는 공격적 경향이 관찰되었으나 통계적으로 유의미하지 않았고, 오직 <실험1>에서만 남성이 여성보다 협동 비율이 유의미하게 낮음을 알 수 있었다($p = 0.0416$). 참가자의 연령이 높을수록 협동 비율이 낮고, 수익도 적으며, 평판 또한 부정적으로 받았으나 통계적으로 유의미하진 않았다.

표 3 <실험1>과 <실험2>의 변수간 상관분석 결과표

<실험 1>	피어슨 상관 계수, N = 60, H0: Rho=0 가정하에서 Prob > |r|								
	참가자의 총 평판	참가자의 협동비율	참가자의 총 긍정적 평판	참가자의 총 부정적 평판	참가자가 부여한 긍정적 평판	참가자가 부여한 부정적 평판	참가자의 수익	참가자의 성별 (여성 0,남성 1)	참가자의 연령
참가자의 총 평판	1	0.79923 <.0001	0.66325 <.0001	−0.97519 <.0001	0.41858 0.0009	−0.48551 <.0001	0.51292 <.0001	−0.17014 0.1937	−0.06159 0.6401
참가자의 협동 비율	0.79923 <.0001	1	0.40675 0.0013	−0.81588 <.0001	0.4842 <.0001	−0.44823 0.0003	0.53255 <.0001	−0.26395 0.0416	−0.08522 0.5174
참가자의 총 긍정적 평판	0.66325 <.0001	0.40675 0.0013	1	−0.48113 <.0001	0.16123 0.2184	−0.23301 0.0732	0.39069 0.002	−0.14845 0.2576	−0.14739 0.2611
참가자의 총 부정적 평판	−0.97519 <.0001	−0.81588 <.0001	−0.48113 <.0001	1	−0.44263 0.0004	0.49979 <.0001	−0.48526 <.0001	0.15539 0.2358	0.02855 0.8285
참가자가 부여한 긍정적 평판	0.41858 0.0009	0.4842 <.0001	0.16123 0.2184	−0.44263 0.0004	1	−0.75426 <.0001	0.46757 0.0002	−0.14942 0.2545	0.04713 0.7206
참가자가 부여한 부정적 평판	−0.48551 <.0001	−0.44823 0.0003	−0.23301 0.0732	0.49979 <.0001	−0.75426 <.0001	1	−0.47927 0.0001	0.16484 0.2082	−0.20048 0.1246
참가자의 수익	0.51292 <.0001	0.53255 <.0001	0.39069 0.002	−0.48526 <.0001	0.46757 0.0002	−0.47927 0.0001	1	−0.23563 0.0699	−0.11564 0.379
참가자의 성별 (여성=0,남성=1)	−0.17014 0.1937	−0.26395 0.0416	−0.14845 0.2576	0.15539 0.2358	−0.14942 0.2545	0.16484 0.2082	−0.23563 0.0699	1	0.19542 0.1346
참가자의 연령	−0.06159 0.6401	−0.08522 0.5174	−0.14739 0.2611	0.02855 0.8285	0.04713 0.7206	−0.20048 0.1246	−0.11564 0.379	0.19542 0.1346	1

<실험 2>	피어슨 상관 계수, N = 60, H0: Rho=0 가정하에서 Prob > |r|								
	참가자의 총 평판	참가자의 협동비율	참가자의 총 긍정적 평판	참가자의 총 부정적 평판	참가자가 부여한 긍정적 평판	참가자가 부여한 부정적 평판	참가자의 수익	참가자의 성별 (여성 0,남성 1)	참가자의 연령
참가자의 총 평판	1	0.73941 <.0001	0.48279 <.0001	−0.97877 <.0001	0.24825 0.0558	−0.27098 0.0362	0.50165 <.0001	−0.21292 0.1024	−0.10755 0.4134
참가자의 협동 비율	0.73941 <.0001	1	0.40842 0.0012	−0.71167 <.0001	0.28794 0.0257	−0.31588 0.0139	0.51165 <.0001	−0.17856 0.1722	−0.11114 0.3979
참가자의 총 긍정적 평판	0.48279 <.0001	0.40842 0.0012	1	−0.29304 0.0231	0.04827 0.7142	0.00323 0.9805	0.02501 0.8496	−0.11285 0.3906	−0.21291 0.1024
참가자의 총 부정적 평판	−0.97877 <.0001	−0.71167 <.0001	−0.29304 0.0231	1	−0.25973 0.0451	0.29661 0.0214	−0.54183 <.0001	0.20604 0.1142	0.06759 0.6079
참가자가 부여한 긍정적 평판	0.24825 0.0558	0.28794 0.0257	0.04827 0.7142	−0.25973 0.0451	1	−0.84581 <.0001	0.36862 0.0038	−0.13293 0.3113	−0.07494 0.5693
참가자가 부여한 부정적 평판	−0.27098 0.0362	−0.31588 0.0139	0.00323 0.9805	0.29661 0.0214	−0.84581 <.0001	1	−0.31092 0.0156	0.20713 0.1123	0.05085 0.6996
참가자의 수익	0.50165 <.0001	0.51165 <.0001	0.02501 0.8496	−0.54183 <.0001	0.36862 0.0038	−0.31092 0.0156	1	−0.22876 0.0787	−0.14266 0.2769
참가자의 성별 (여성=0,남성=1)	−0.21292 0.1024	−0.17856 0.1722	−0.11285 0.3906	0.20604 0.1142	−0.13293 0.3113	0.20713 0.1123	−0.22876 0.0787	1	0.19542 0.1346
참가자의 연령	−0.10755 0.4134	−0.11114 0.3979	−0.21291 0.1024	0.06759 0.6079	−0.07494 0.5693	0.05085 0.6996	−0.14266 0.2769	0.19542 0.1346	1

2. 사람들은 부정적 평판 부여 편향을 가지고 있는가?

참가자들간 가진 자원이 비슷한 경우 상대의 이타적 행동에 대한 긍정적 평판보다 이기적 행동에 대한 부정적 평판이 더 강하다는 부정적 평판 평가 편향성을 검증하기 위하여 <실험1>의 자료를 사용하여 협동 횟수와 배신 횟수가 최종 평판에 미치는 영향을 분석하였다. 최종 평판 점수는 긍정적인 평판 점수와 부정적인 평판 점수의 단순 합이기 때문에 각각을 분리하여 이타적 행동이 긍정적 평판이 미치는 영향과 이기적 행동이 부정적 평판에 미치는 영향을 따로 살펴볼 필요가 있다. 따라서 SAS 프로그램 ver 9.1.3을 사용하여 협동 횟수와 긍정적인 평판 점수의 단순 회귀 분석(Single Regression Analysis)과 배신 비율과 부정적 평판 점수의 단순 회귀 분석(Single Regression Analysis)을 각각 유의수준 5%에서 실시함으로써 협동과 배신 행동에 따라 변화하는 긍정적 평판과 부정적 평판의 상승폭을 추적하였다. 그 뒤 등분산 검증을 한 뒤 이표본 t-test를 통하여 두 값 N과 P의 차이의 유의미성을 검증하였다.

각각의 단순 회귀식은 $Y = \alpha + \beta * x$ 의 형태로 표현되며 종속변수 Y 는 긍정적 혹은 부정적 평판이고 x 는 협동 횟수 혹은 배신 횟수 임으로 이 때의 β 값을 협동 행위에 대한 긍정적 평판의 상승폭 P 와 배신 행위에 대한 부정적 평판의 상승폭 N 으로 해석할 수 있다. 이를 통해 협동 횟수와 긍정적 평판의 단순 회귀 식을 변형한 식 $R_p = 1.396 + 1.485 \cdot r_p$ (F 1,59 = 11.5, p = 0.0013)에서 나타난 P 값은 1.485이며, 배신 횟수와 부정적 평판의 단순 회귀 식을 변형한 식 $R_n = 10.341 + 10.071 \cdot r_n$ (F 1,59 = 115.48, p < 0.0001)을 통하여 N값은 10.071임을 도출할 수 있다(<표 4>). 또한 두 값 N과 P의 차이의 유의미성을 검증 결과, 이타적 행동에 대한 긍정적 평판 점수와 이기적 행동에 대한 부정적 평판 점수가 등분산을 이루며(F= 5.08 p < 0.0001), 이타적 행동에 대한 긍정적 평판 상승폭보다 이기적 행동에 대한 부정적 평판 상승폭이 더 큼을 알 수 있었다(t = 15.48, p< 0.0001). 따라서 실험의 참가자들은 부정적 평판 평가 편향을 가지고 있다고 할 수 있다.

표 4 각 행동에 따른 평판의 단순 회귀식 및 상승폭P와 상승폭N

이타적 행동에 따른 긍정적 평판의 단순 회귀식: $\mathbf{R_p = P \cdot r_p + e}$	**상승폭 P**
$R_p = 1.485 \cdot r_p + 1.396$ (F 1,59 = 11.5, p = 0.0013)	1.485
이기적 행동에 따른 부정적 평판의 단순 회귀식: $\mathbf{R_n = N \cdot r_n + e}$	**상승폭 N**
$R_n = 10.071 \cdot r_n + 10.341$ (F 1,59 = 115.48, p < 0.0001)	10.071

R_p : 긍정적 평판, P : 긍정적 평판 상승폭, r_p : 이타적 행동 (1회)

R_n : 부정적 평판, N : 부정적 평판 상승폭, r_n : 이기적 행동 (1회)

3. 가설에서 설정한 각 요인들에 따라 참가자들은 평판 점수를 어떻게 부여하는가?

연구 가설에서 상대방의 협동 여부뿐만 아니라 자신의 협동 여부, 게임의 성공 여부, 상대적 자원차이가 평판 평가에 영향을 미칠 것이라 설정하였다. 이를 알아보기 위하여 <실험1>과 <실험2>에서 참가자들이 상대방에게 부여한 긍정적 평판과 부정적 평판을 SAS 프로그램 ver 9.1.3을 사용하여 분석하였다. 각 요인에 따라 구분한 점수를 유의수준 5%에서 일원 배치 분산 분석 (1-way ANOVA analysis)을 통해 차이가 있음을 증명한 후 Tukey의 표준화 검정을 통한 단순 평균비교를 실시하였다.

긍정적 평판 점수는 상대방이 협동할 경우가 배신할 때 보다 평균 0.2237, 자신이 협동하였을 때가 자신이 배신하였을 때보다 평균 0.2871점, 게임이 성공하였을 때가 실패하였을 때보다 평균 0.3692점 높게 나타났다. 또한 상대방의 자원이 자신보다 적거나 자신과 같을 때가 상대방의 자원이 자신보다 많을 때보다 평균 0.3664점 높았다(<표 5>). 즉 참가자들은 상대방이 협동할 경우, 자신이 협동할 경우, 게임이 성공할 경우, 자신이 가진 자원이 상대방보다 많거나 상대방과 같은 경우에서 상대방에게 긍정적 평판 점수를 더 많이 주었다.

부정적 평판 점수는 상대방의 협동 여부에 따라 통계적으로 유의미한 차이가 나타나지 않았다. 그러나 자신이 배신하였을 때가 협동하였을 때 보다 평균 0.3739점, 게임이 실패하였을 때가 성공하였을 때보다 평균 0.4437점, 상대방의 자원이 자신보다 많을 때가 상대방의 자원이 자신보다 적거나 같을 때보다 평균 0.9543점 더 크게 나타났다(<표 6>). 즉 참가자들은 자신이 배신한 경우, 게임이 실패한 경우, 상대방의 자원이 자신보다 많은 경우에서 상대방에게 부정적 평판 점수를 더 많이 주었다.

표 5 각 상황에 따라 상대방에게 부여한 긍정적 평판 점수의 평균 비교

TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(상대방의 협동) - (상대방의 배신)	0.2237	0.06834	0.37897	*
(자신의 협동) - (자신의 배신)	0.2871	0.13175	0.44252	*
(게임의 성공) - (게임의 실패)	0.3692	0.21257	0.52575	*
(자원: 상대방 ≤ 자신) - (자원: 상대방 > 자신)	0.3664	0.21162	0.52113	*

*는 유의 수준 5%에서 유의미함을 표시함

표 6 각 상황에 따라 상대방에게 부여한 부정적 평판 점수의 평균 비교

TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(상대방의 협동) - (상대방의 배신)	0.0073	−0.27219	0.28696	
(자신의 협동) - (자신의 배신)	−0.3739	0.6536	0.0942	*
(게임의 성공) - (개임의 실패)	−0.4437	0.7255	0.1618	*
(자원: 상대방 ≤ 자신) - (자원: 상대방 > 자신)	−0.9543	1.2329	0.6758	*

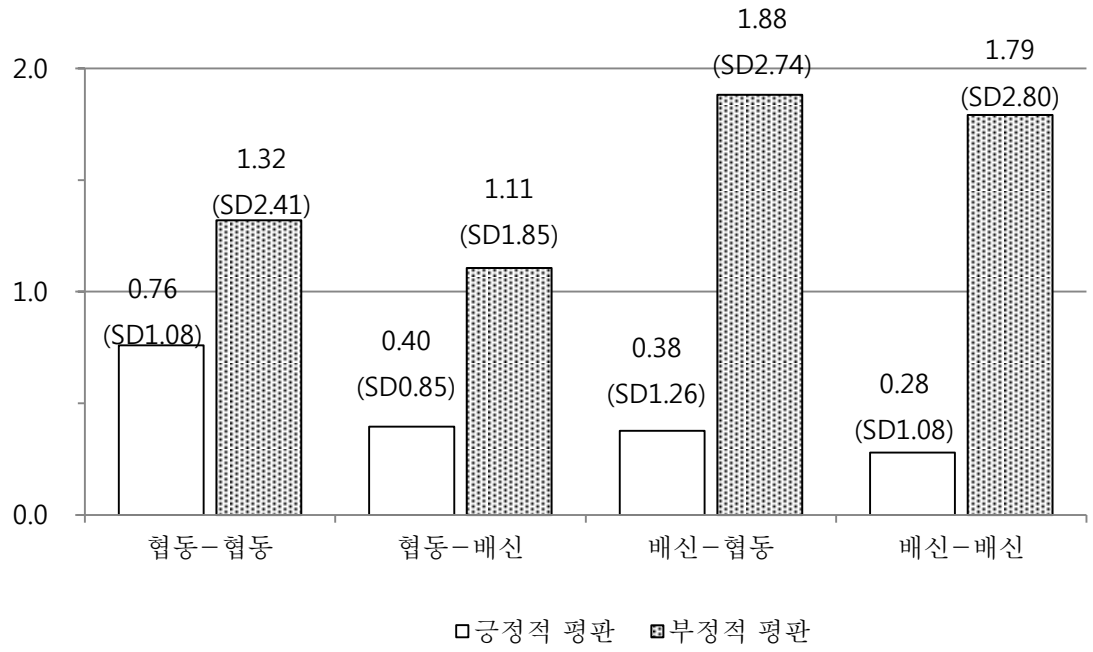
*는 유의 수준 5%에서 유의미함을 표시함

4. 자신과 상대방의 협동 여부에 따라 평판 점수 부여 경향은 다르게 나타나는가?

참가자들간 가진 자원이 비슷한 경우 자신의 협동 여부와 상대방의 협동 여부에 따른 상대방에대한 평판 경향을 알아보기 위하여 <실험1>의 자료를 사용하였다. 자신이 협동하였을 때 상대방도 협동하는 상황(협동-협동), 자신이 협동하였을 때 상대방은 배신한 상황(협동-배신), 자신이 배신하였을 때 상대방이 협동한 상황(배신-협동), 자신이 배신하였을 때 상대방도 배신한 상황(배신-배신)의 총 4가지 상황에서 나타나는 긍정적 평판의 평균과 부정적 평판을 각각 구분하여 살펴보았다. 이 때 4가지 상황을 하나의 요인으로 설정하고 긍정적 평판 점수와 부정적 평판 점수 각각에 대하여 일원배치 분산 분석 (1-way ANOVA analysis)을 통해 유의수준 5%에서 차이가 있음을 증명한 후 Tukey의 표준화 검정을 이용한 다중비교 (multiple comparison)를 실시하여 어느 상황 간 차이가 존재하는지 SAS 프로그램 ver 9.1.3으로 분석하였다.

자신이 협동하였을 때 협동한 상대방에게 부여하는 긍정적 평판의 사례 수는 935개, 자신이 협동하였을 때 배신한 상대방에게 부여하는 긍정적 평판의 사례 수는 205개였으며 자신이 배신하였을 때 협동한 상대방에게 부여하는 긍정적 평판의 사례 수는 204개, 자신이 배신하였을 때 배신한 상대방에게 부여하는 긍정적 평판의 사례 수는 336개로 나타났다. 자신과 상대방의 협동 여부에 따른 긍정적 평판과 부정적 평판의 통계값은 다음 <그림 8>과 같다.

그림 8 자신의 협동 여부와 상대방의 협동 여부에 따라 상대방에게 부여한 평판 점수



자신의 협동 여부와 타인의 협동 여부에 따라 구분한 협동−협동 상황(i=1), 협동−배신 상황(i=2), 배신−협동 상황(i=3), 배신−배신 상황(i=4)의 총 4 가지 상황에서 배신한 참가자는 협력한 참가자에 비해 다른 배신한 참가자에게 부정적 평판을 가하지 않고 협력한 참가자에게 긍정적 평판을 할 가능성이 크고 협력한 참가자는 다른 협력한 참가자에게 긍정적 평판을 주기보다 무임승차자에게 부정적 평판을 가할 가능성이 더 클 것이란 예상과 다르게 나타났다. 긍정적 평판 $\mathrm{RP_i}(\mathrm{i}=1,2,3,4)$ 중 자신과 상대방 모두 협동한 상황인 협동−협동 상황에서 상대에게 부여한 긍정적 평판만이 다른 상황들보다 더 크게 나타남으로써 통계적으로 유의미한 차이를 보였다($\mathrm{RP_1} \neq \mathrm{RP_2}, \mathrm{RP_3}, \mathrm{RP_4}$ <표 7>). 한편 부정적 평판 $\mathrm{RN_i}(\mathrm{i}=1,2,3,4)$ 의 경우 자신의 협동 여부에 따라 차이를 보였는데 자신이 협동하였을 때보다 배신하였을 때 상대방의 협동 유무에 상관없이 상대방에 대한 부정적 평판이 통계적으로 유의미하게 크게 나타났다($\mathrm{RN_1}, \mathrm{RN_2} \neq \mathrm{RN_3}, \mathrm{RN_4}$ <표 8>). 즉 참가자들은 자신과 상대방 모두 협동한 경우에만 상대방을 더 긍정적으로 평판하였고, 자신이 배신한 경우에 협동한 경우보다 상대방을 더 부정적으로 평판하였음을 알 수 있다.

표 7 자신과 상대방의 협동 여부에 따라 상대방에게 부여한 긍정적 평판 점수 비교(i=1,2,3,4)

TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신 협동, 상대방 협동) – (자신 협동, 상대방 배신)	0.36424	0.15012	0.57835	*
(자신 협동, 상대방 협동) – (자신 배신, 상대방 협동)	0.38191	0.16736	0.59645	*
(자신 협동, 상대방 협동) – (자신 배신, 상대방 배신)	0.47960	0.30300	0.65619	*
(자신 협동, 상대방 배신) – (자신 배신, 상대방 협동)	0.01767	−0.25689	0.29223	
(자신 협동, 상대방 배신) – (자신 배신, 상대방 배신)	0.11536	−0.13069	0.36141	
(자신 배신, 상대방 협동) – (자신 배신, 상대방 배신)	0.09769	−0.14874	0.34411	

*는 유의 수준 5% 에서 유의미함을 표시함

표 8 자신과 상대방의 협동 여부에 따라 상대방에게 부여한 부정적 평판 점수 비교(i=1,2,3,4)

TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신 협동, 상대방 협동) – (자신 협동, 상대방 배신)	0.2135	−0.2778	0.7049	
(자신 협동, 상대방 협동) – (자신 배신, 상대방 협동)	−0.5615	−1.0538	−0.0691	*
(자신 협동, 상대방 협동) – (자신 배신, 상대방 배신)	−0.4708	−0.8761	−0.0656	*
(자신 협동, 상대방 배신) – (자신 배신, 상대방 협동)	−0.7750	−1.4051	−0.1449	*
(자신 협동, 상대방 배신) – (자신 배신, 상대방 배신)	−0.6843	−1.2490	−0.1197	*
(자신 배신, 상대방 협동) – (자신 배신, 상대방 배신)	0.0907	−0.4748	0.6562	

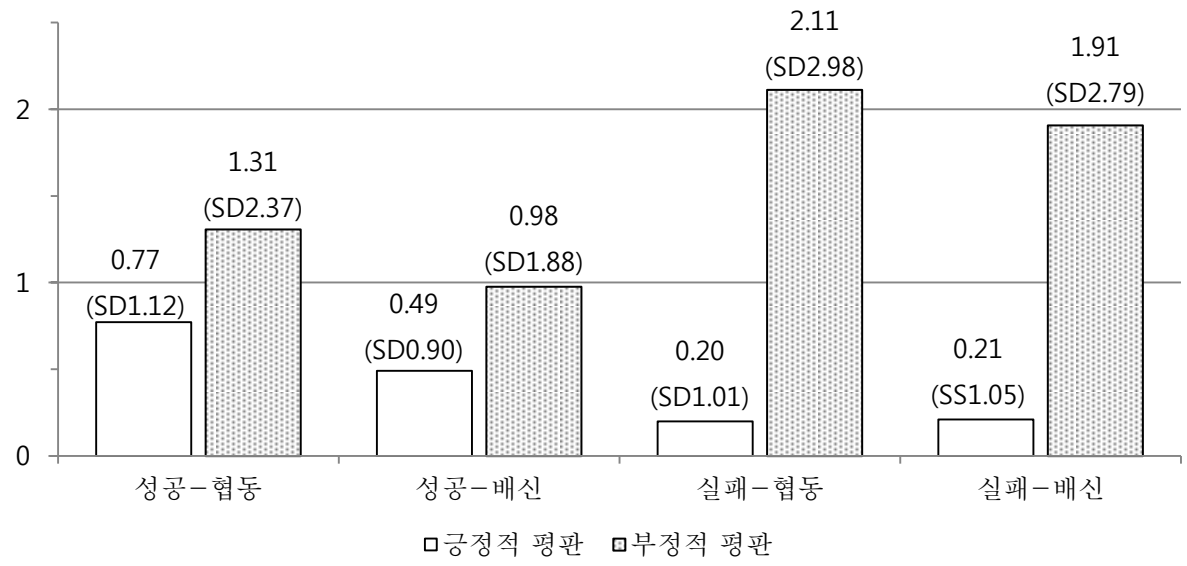
*는 유의 수준 5% 에서 유의미함을 표시함

5. 게임의 성공 여부와 상대방의 협동 여부에 따라 평판 평가 경향은 다르게 나타나는가?

참가자들간 가진 자원이 비슷한 경우 게임의 성공 여부와 상대방의 협동 여부에 따른 평판 경향을 알아보기 위하여 <실험1>의 자료를 사용하였다. 게임이 성공하였을 때 상대방도 협동하는 상황(성공-협동), 게임이 성공하였을 때 상대방은 배신한 상황(성공-배신), 게임이 실패하였을 때 상대방이 협동한 상황(실패-협동), 게임이 실패하였을 때 상대방도 배신한 상황(실패-배신)의 총 4가지 상황에서 나타나는 긍정적 평판의 평균과 부정적 평판을 각각 구분하여 살펴보았다. 이 때 4가지 상황을 하나의 요인으로 설정하고 긍정적 평판 점수와 부정적 평판 점수 각각에 대하여 일원 배치 분산 분석 (1-way ANOVA analysis)을 통해 유의수준 5%에서 차이가 있음을 증명한 후 Tukey의 표준화 검정을 이용한 다중비교(multiple comparison)를 실시하여 어느 상황 간의 차이가 존재하는지 SAS 프로그램 ver 9.1.3으르 분석하였다.

게임이 성공하였을 때 협동한 상대방에게 부여하는 긍정적 평판의 사례 수는 978개, 게임이 성공하였을 때 배신한 상대방에게 부여하는 긍정적 평판의 사례 수는 218개였으며, 게임이 실패하였을 때 협동한 상대방에게 부여하는 긍정적 평판의 사례 수는 161개, 게임이 실패하였을 때 배신한 상대방에게 부여하는 긍정적 평판의 사례 수는 323개였다. 게임의 성공 여부와 상대방의 협동 여부에 따른 긍정적 평판과 부정적 평판의 통계값은 <그림 9>와 같다.

그림 9 게임의 성공 여부와 상대방의 협동 여부에 따라 상대방에게 부여한 평판 점수



게임의 성공 여부와 타인의 협동 여부에 따라 분류한 성공-협동 상황(i=1), 성공-배신 상황(i=2), 실패-협동 상황(i=3), 실패-배신 상황(i=4)의 총 4 가지 상황에서 나타나는 긍정적 평판 $\mathrm{RP_i}(i = 1,2,3,4)$ 의 분석 결과, 참가자들은 게임이 성공하고 상대방이 협동하였을 때 가장 긍정적으로 평판하였으며, 게임이 성공하면 상대방이 배신을 하더라도 긍정적인 평판을 하였음을 알 수 있었다. 반면 게임이 실패하였을 경우 상대방의 협동여부에 따라서는 통계적으로 유의미한 차이가 발견되지 않았다 ($\mathrm{RP_1} \neq \mathrm{RP_2}, \mathrm{RP_3}, \mathrm{RP_4}$ 이고 $\mathrm{RP_2} \neq \mathrm{RP_3}, \mathrm{RP_4}$ <표 9>). 다시 말하자면 게임이 실패하였을 경우 참가자들은 상대방의 협동 여부에 상관없이 상대방에 대한 긍정적 평판을 많이 하지 않는 반면, 게임이 성공하였을 경우만 긍정적으로 평판하였다. 특히 게임이 성공했을 경우 참가자들은 상대방이 배신했을 때보다 협동하였을 때 가장 상대방에게 긍정적 평판을 많이 주었다. 또한 총 4 가지 상황에 따른 부정적 평판 $\mathrm{RN_i}(i = 1,2,3,4)$ 은 게임의 성공 여부에 따라 차이를 보였는데 상대방의 협동 유무에 상관없이 게임이 성공했을 때는 게임이 실패하였을 때보다 부정적 평판이 통계적으로 유의미하게 작게 나타났다($\mathrm{RN_1}, \mathrm{RN_2} \neq \mathrm{RN_3}, \mathrm{RN_4}$ <표 10>). 즉 참가자들은 게임이 실패하였을 경우 게임이 성공하였을 때보다 상대방에게 더 부정적인 평판을 많이 한다고 할 수 있다.

표 9 게임의 성공과 상대방의 협동 여부에 따라 상대방에게 부여한 긍정적 평판 점수 비교(i=1,2,3,4) TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(게임 성공, 상대방 협동) − (게임 성공, 상대방 배신)	0.28116	0.07479	0.48752	*
(게임 성공, 상대방 협동) − (게임 실패, 상대방 협동)	0.57323	0.33889	0.80757	*
(게임 성공, 상대방 협동) − (게임 실패, 상대방 배신)	0.56146	0.38464	0.73828	*
(게임 성공, 상대방 배신) − (게임 실패, 상대방 협동)	0.29207	0.00575	0.57838	*
(게임 성공, 상대방 배신) − (게임 실패, 상대방 배신)	0.28030	0.03879	0.52181	*
(게임 실패, 상대방 협동) − (게임 실패, 상대방 배신)	−0.01177	−0.27758	0.25404	

*는 유의 수준 5% 에서 유의미함을 표시함

표 10 게임의 성공과 상대방의 협동 여부에 따라 상대방에게 부여한 부정적 평판 점수 비교(i=1,2,3,4) TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(게임 성공, 상대방 협동) − (게임 성공, 상대방 배신)	0.3307	−0.1444	0.8058	
(게임 성공, 상대방 협동) − (게임 실패, 상대방 협동)	−0.8040	−1.3436	−0.2645	*
(게임 성공, 상대방 협동) − (게임 실패, 상대방 배신)	−0.5993	−1.0065	−0.1922	*
(게임 성공, 상대방 배신) − (게임 실패, 상대방 협동)	−1.1347	−1.7939	−0.4755	*
(게임 성공, 상대방 배신) − (게임 실패, 상대방 배신)	−0.9301	−1.4861	−0.3740	*
(게임 실패, 상대방 협동) − (게임 실패, 상대방 배신)	0.2047	−0.4073	0.8167	

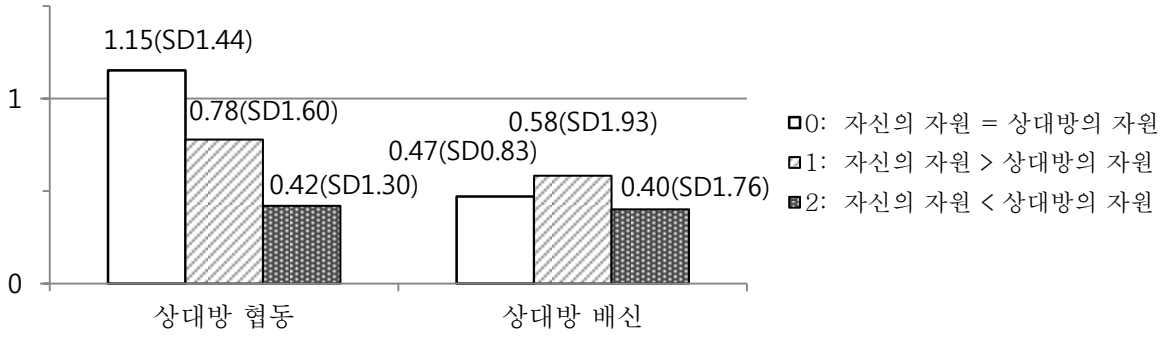
*는 유의 수준 5% 에서 유의미함을 표시함

6. 자원의 상대적 차이와 상대방의 협동 여부에 따라 평판 평가 경향은 다르게 나타나는가?

참가자간 가지고 있는 자원이 비대칭적일 경우 자원의 상대적 차이에 따라 상대방에 대한 평판 경향이 다르게 나타나는지 살펴보기 위해 <실험2>의 결과를 사용하였다. 자신의 자원과 상대방의 자원이 같은 상황(d=0), 자신의 자원이 상대방의 자원보다 적은 상황(d=1), 자신의 자원이 상대방의 자원보다 많은 상황(d=2), 의 총 3가지 상황에서 상대방의 협동 여부에 따라 나타나는 긍정적 평판과 부정적 평판을 각각 구분하여 살펴보았다. 이 때 3가지 상황을 하나의 요인으로 설정하고 상대방이 협동했을 때와 배신했을 때의 긍정적 평판 점수, 상대방이 협동했을 때와 배신했을 때의 부정적 평판 점수에서 각각 유의수준 5%에서 일원 배치 분산 분석 (1-way ANOVA analysis)을 통해 차이가 있음을 증명한 후, Tukey의 표준화 검정을 이용한 다중비교(multiple comparison)를 실시하여 어느 상황 간 차이가 존재하는지 SAS 프로그램 ver 9.1.3으로 분석하였다.

상대방이 협동하였을 때 그 상대가 자신보다 자원이 많을 경우의 사례 수는 395개, 자신보다 자원이 적을 경우의 사례 수는 380개, 자신과 가진 자원이 같은 경우의 사례 수는 211개였다. 상대방이 배신하였을 때 그 상대가 자신보다 자원이 많을 경우의 사례 수는 317개, 자신보다 자원이 적은 경우의 사례 수는 326개, 자신과 상대방의 자원이 같은 경우의 사례수는 51개였다. 상대적 자원차이와 상대방의 협동 여부에 따른 긍정적 평판의 통계값은 <그림 10>과 같다.

그림 10 상대방의 협동 여부와 상대적 자원 차이에 따라 상대방에게 부여한 긍정적 평판 점수



상대방이 협동하는 상황(i=1)과 상대방이 배신하는 상황(i=0)을 나누어 살펴보았을 때, 자원의 상대적인 차이에 따른 긍정적 평판의 경우 통계적으로 유의미한 차이는 상대방이 협동하는 상황에서만 나타났다. 상대방이 협동한 경우 나타나는 긍정적 평판 $\mathrm{RP_{id}}(i=1, d=0,1,2)$ 은 자신과 상대방의 자원 차이에 따라 모두 통계적으로 유의미한 차이가 있었다($\mathrm{RP_{10}} \neq \mathrm{RP_{11}} \neq \mathrm{RP_{12}}$<표 11>). 즉 참가자들은 상대방이 자원을 많이 가진 경우 자신과 자원 차이가 없거나 자신보다 적게 가진 사람보다 긍정적 평판을 덜함으로써 많은 자원을 가진 상태가 부정적 평판 평가 편향을 강화시킨다는 연구가설을 지지하였다. 그러나 자신보다 자원이 적은 사람이 협동하였을 때 더 긍정적으로 평판하진 않았다. 오히려 자신과 자원의 차이가 없는 상대에게 가장 긍정적으로 평판하는 경향이 강하게 나타났다.

반면 상대방이 배신하였음에도 불구하고 부여하는 긍정적 평판 $\mathrm{RP_{id}}(i=0, d=0,1,2)$ 점수는 그 양이 매우 작게 나타났다. 단순 통계량을 살펴보면 각각의 평균의 순위가 $\mathrm{RP_{01}}$, $\mathrm{RP_{00}}$, $\mathrm{RP_{02}}$ 순으로, 상대방의 자원이 자신보다 적을수록 긍정적 평판의 양이 증가하고, 자신보다 상대적으로 자원이 많을수록 긍정적 평판이 감소하는 연구 가설의 내용과 일치하였으나 통계적으로 유의미한 차이는 발견되지 않았다(<표 12>).

표 11 상대방 협동 시(i=1) 상대적 자원 차이(d=0,1,2)에 따라 상대방에게 부여한 긍정적 평판 점수 비교 TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신의 자원=상대방의 자원) − (자신의 자원>상대방의 자원)	0.3727	0.0805	0.6649	*
(자신의 자원=상대방의 자원) − (자신의 자원<상대방의 자원)	0.7314	0.4412	1.0216	*
(자신의 자원>상대방의 자원) − (자신의 자원<상대방의 자원)	0.3587	0.1141	0.6033	*

*는 유의 수준 5% 에서 유의미함을 표시함

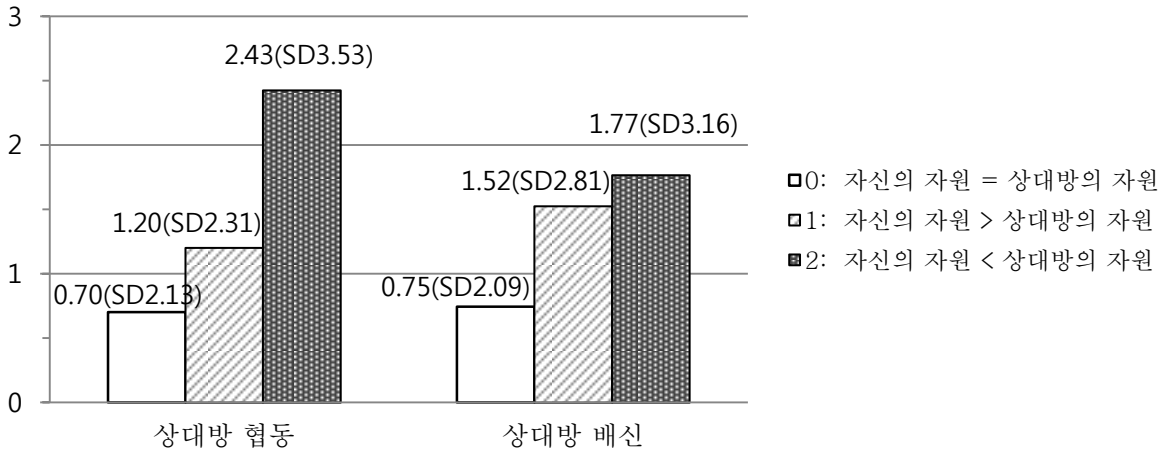
표 12 상대방 배신 시(i=0) 상대적 자원 차이(d=0,1,2)에 따라 상대방에게 부여한 긍정적 평판 점수 비교 TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신의 자원=상대방의 자원) − (자신의 자원>상대방의 자원)	−0.1122	−0.7462	0.5217	
(자신의 자원=상대방의 자원) − (자신의 자원<상대방의 자원)	0.0700	−0.5652	0.7051	
(자신의 자원>상대방의 자원) − (자신의 자원<상대방의 자원)	0.1822	−0.1499	0.5143	

*는 유의 수준 5% 에서 유의미함을 표시함

상대적 자원차이와 상대방의 협동 여부에 따른 부정적 평판의 통계값은 <그림 11>과 같다.

그림 11 상대방의 협동 여부와 상대적 자원 차이에 따라 상대방에게 부여한 부정적 평판 점수



상대방이 협동하는 상황(i=1)과 상대방이 배신하는 상황(i=0)을 나누어 살펴보았을 때, 부정적 평판의 경우에서도 통계적으로 유의미한 차이는 상대방이 협동하는 상황에서만 나타났다. 상대방이 협동하였음에도 불구하고 부정적인 평판을 주는 부당한 평판주기이다. 부당한 부정적 평판 $\mathrm{RN_{id}}(i=1, d=0,1,2)$ 는 자신보다 상대방의 자원이 많은 경우에 자신보다 상대방이 자원이 적거나 자신과 같은 경우보다 더 크게 나타남으로써 연구 가설의 내용과 비슷하였다($\mathrm{RN_{12}} \neq \mathrm{RN_{11}}, \mathrm{RN_{10}}$<표 13>). 즉 상대방이 협동하였음에도 불구하고 자신보다 자원이 많은 경우 부정적 평판을 많이 내림으로써 참가자들은 질투의 감정을 표출하고 상대적인 자원 차이를 줄이기 위해 노력했다고 볼 수 있다.

반면 상대방이 배신한 경우 나타나는 부정적 평판 $\mathrm{RN_{id}}(i=0, d=0,1,2)$ 은 각각의 평균의 순위가 $\mathrm{RN_{00}}$, $\mathrm{RN_{01}}$, $\mathrm{RN_{02}}$ 순으로 나타났다. 배신 행동을 하였어도 자신과 자원을 비슷하게 가진 경우 부정적인 평판을 가장 적게 함으로써 자신보다 자원이 적은 상대방에게 가장 덜 부정적인 평판을 할 것이라는 연구가설을 지지하지는 않았다.

그러나 연구 가설에서 설정한 내용과 동일하게 자원을 많이 가진 상태는 이기적 행동을 하였을 때 다른 참가자들로부터 가장 부정적 평판을 많이 받게끔 만들었다. 이는 전반적으로 자원을 많이 가진 상태가 부정적 평판 평가 편향을 강화시킬 것이라는 연구 가설의 내용과 비슷한 측면이 있었으나 통계적으로 유의미한 차이는 발견되지 않았다(<표 14>)

표 13 상대방 협동 시(i=1) 상대적 자원 차이(d=0,1,2)에 따라 상대방에게 부여한 부정적 평판 점수 비교 TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신의 자원=상대방의 자원) − (자신의 자원>상대방의 자원)	−0.5012	−1.0720	0.0696	
(자신의 자원=상대방의 자원) − (자신의 자원<상대방의 자원)	−1.7239	−2.2908	−1.157	*
(자신의 자원>상대방의 자원) − (자신의 자원<상대방의 자원)	−1.2227	−1.7004	−0.745	*

*는 유의 수준 5% 에서 유의미함을 표시함

표 14 상대방 배신 시(i=0) 상대적 자원 차이(d=0,1,2)에 따라 상대방에게 부여한 부정적 평판 점수 비교 TUKEY'S STUDENTIZED RANGE (HSD)TEST

비교	평균차	95% 신뢰구간		
(자신의 자원=상대방의 자원) − (자신의 자원>상대방의 자원)	−0.7794	−1.8169	0.2580	
(자신의 자원=상대방의 자원) − (자신의 자원<상대방의 자원)	−1.0215	−2.0609	0.0180	
(자신의 자원>상대방의 자원) − (자신의 자원<상대방의 자원)	−0.2420	−0.7855	0.3014	

*는 유의 수준 5% 에서 유의미함을 표시함

7. 평판 평가에 각 요인들이 미치는 영향력은 어떻게 나타나는가?

상대방의 협동 여부, 자신의 협동 여부, 게임의 성공 여부, 상대적 자원 차이가 상대방에 대한긍정적 평판 평가와 부정적 평판 평가에 미치는 영향력의 크기를 살펴보기 위하여 <실험1>과 <실험2>의 자료를 사용하여 SAS 프로그램 ver 9.1.3으로 다중회귀분석(multiple regression analysis)을 실시하였다. 이 때 긍정적 평판과 부정적 평판 각각의 회귀식은 $Y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \varepsilon$ 로 표현되며 종속변수 Y 는 긍정적 혹은 부정적 평판이고 x 는 각 요인 변수이며 이 때의 β 값은 각 요인이 긍정적 혹은 부정적 평판에 미치는 영향력이라 해석할 수 있다. 이 때 각 요인들의 절대값 β 를 비교하여 평판 평가에 영향을 미치는 비중을 살펴보았다.

분석 결과 긍정적 평판에 영향을 미치는 각 요인들의 β 값은 <표 15>과 같다. 참가자들은 상대방의 자원이 자신의 자원보다 많은 경우 긍정적 평판을 적게 하였고, 게임이 성공한 경우와 자신이 협동한 경우 더 긍정적으로 평판하였음을 알 수 있었다. 이 때 상대방의 협동 여부와 자신의 협동 여부는 유의수준 5%에서 통계적으로 유의미하지 않게 나타나 상대방에게 부여한 긍정적 평판 점수에 대한 두 요인의 설명력이 낮음을 알 수 있다. 또한 각 요인의 절대값 β 를 비교하여 자원의 상대적 차이, 게임의 성공 여부, 자신의 협동 여부, 상대방 협동 여부 순으로 긍정적 평판에 큰 영향력을 가짐을 알 수 있었다. 상대방이 협동한 경우 긍정적 평판을 적게 한 것으로 나왔으나 다른 요인에 비해 β 의 절대값이 매우 작기 때문에 유의미하지 않다.

표 15 상대방에게 부여한 긍정적 평판 점수에 영향을 미치는 각 요인들의 Beta 값

독립변수	상대방 협동	자신 협동	게임 성공	자원: 상대방>자신
β	−0.0074	0.1812	0.2447 *	−0.3591 ***

***는 유의 수준 0.1% **는 유의수준 1% *는 유의수준 5%에서 유의미함을 표시함

부정적 평판에 영향을 미치는 각 요인들의 β 값은 <표 16>과 같다. 참가자들은 상대방이 배신한 경우, 자신이 배신한 경우, 게임이 실패한 경우, 상대방이 자신보다 자원이 많은 경우에 부정적 평판을 더 많이 하였음을 알 수 있다. 또한 β 의 절대값을 비교하여 자원의 상대적 차이, 자신의 협동 여부, 게임의 성공 여부, 상대방의 협동 여부 순으로 부정적 평판 평가에 큰 영향력을 가짐을 알 수 있었다.

표 16 상대방에게 부여한 부정적 평판 점수에 영향을 미치는 각 요인들의 Beta 값

독립변수	상대방 배신	자신 배신	게임 실패	자원: 상대방>자신
β	0.35361 **	0.43486 *	0.37314 *	0.97812 ***

***는 유의 수준 0.1% **는 유의수준 1% *는 유의수준 5%에서 유의미함을 표시함

Ⅶ. 토론 및 결론

긍정적 평판은 비싼 신호이론의 확장으로써 보다 규모가 큰 집단에 적용시킬 수 있는 장점이 있으며, 부정적 평판은 이타적 처벌과 함께 작동함으로써 가장 효과적으로 무임승차를 견제할 수 있도록 한다. 따라서 평판은 이타적 행동의 이해에 있어서 진화적 관점의 기존 이론들을 완전히 대체하는 것이 아닌 보완하는 요소로 이해하는 것이 바람직하다. 본 논문은 이러한 맥락에서 보상과 처벌에 절대적으로 의존하지 않더라도 효과적으로 협력이 유지될 가능성을 보여주었다. 평판 평가의 규칙이 추가된 공공재 게임에서 보상이나 이타적 처벌이 없음에도 불구하고 무임승차자는 협력자보다 큰 이익을 얻을 수 없었다. 오히려 협동 비율이 높은 참가자들은 긍정적 평판뿐만 아니라 수익 또한 높았다. 협동 비율이 높은 참가자는 상대방에게 긍정적 평판을 얻을 수 있었고, 이는 반복되는 게임에서 좋은 팀원을 만날 확률을 높였다. 결과적으로 협력비율이 높은 참가자들은 협동을 통해 발생되는 공공의 이익을 나눠가지는 방법으로 개인의 수익을 높일 수 있었다. 반대로 어떤 참가자가 배신을 통해 한 라운드에서 협력자보다 이익을 얻어도 다른 참가자들에 의해 나쁜 평판을 받게 된다. 이에 따라 협력을 잘하는 평판이 좋은 사람을 만날 기회가 줄어들고, 결국 공공의 협력을 통해 발생할 이익을 접할 수 없기 때문에 장기적으로는 낮은 수익을 얻는 것이다.

부정적 평판 평가 편향이 존재할 것이라는 예상은 실험 결과와 일치하였다. 이타적 행동에 대한 긍정적 평판의 상승폭과 이기적 행동에 대한 부정적 평판의 상승폭을 도출하여 비교한 결과 부정적 평판의 상승폭이 더 크게 나타났다. 이는 평판이 보상과 처벌의 효과를 가지고 있다고 볼 때 반복되는 공공재 게임에서 참가자들은 보상보다 처벌을 점차 선호할 뿐 아니라 처벌하려는 동기도 보상하려는 동기보다 더 강하다는 기존의 연구(Gürerk et al, 2004; Price, et al., 2002)와 맥락을 같이 한다고 할 수 있다. 이는 다른 참가자에게 추가적인 이익이 될 수 있는 긍정적 평판을 주기보다는 부정적 평판을 가하여 무임승차자에게 불이익을 주는 것이 상대적인 손실이나 박탈감 없이

협동을 유지하는데 더 유리하기 때문이다.

한편 실험 전반에 걸쳐 부당한 평판하기가 부정적 평판을 중심으로 많이 관찰되었다. 부당한 처벌은 경쟁심에 의해 전략적으로 사용되는데(김상인, 2006) 본 연구의 실험에서도 같은 동기가 작동하였다. 즉 참가자들이 협동을 하더라도 불공정하게 부정적 평판을 받는 가장 큰 이유는 평판 경쟁 때문이다. 평판에 따라 새로운 팀 배정이 이루어지기 때문에 참가자들은 평판 경쟁을 통해 평판이 좋은 다른 참가자를 만나 게임을 성공시켜 이점을 취하려는 동기가 강하였다. 따라서 협동한 상대방에게도 부정적 평판을 내리거나 긍정적 평판을 소극적으로 사용함으로써, 자신이 더 좋은 팀을 만날 확률을 높이려는 경향이 강하였다. 또한 남성은 여성보다 협동 비율이 낮고 상대방에게 부정적 평판을 더 많이 가하는 공격적인 경쟁 성향을 보였다. 이는 남성이 여성보다 위험을 감수하려는 경향이 강하고 경쟁적이며 공격적이라는 기존의 연구결과와 일치한다(Wilson & Daly, 1985).

기존의 평판을 사용한 연구들은 평판의 대상인 상대방의 협동 유무만을 고려하였다는 한계가 있었다. 그러나 실험 결과 상대방의 협동 여부보다 자신의 협동 여부, 게임의 성공 여부, 상대방과 자신의 상대적 자원 차이가 평판 평가 과정에 더 큰 영향을 미칠 수도 있음을 발견하였다. 긍정적 평판의 경우 상대방의 자원이 나보다 적거나 나보다 같을 경우, 게임이 성공한 경우, 자신이 협동한 경우, 상대방이 협동한 경우 순으로 참가자들이 상대방에게 긍정적 평판을 할 가능성이 높으며, 부정적 평판의 경우 상대방의 자원이 자신보다 많을 경우, 자신이 배신한 경우, 게임이 실패한 경우, 상대방이 배신한 경우 순으로 참가자들이 상대방에게 부정적 평판을 할 가능성이 높다는 것을 알 수 있었다.

이어서 위의 각 상황에서 참가자들이 가진 상대방의 협동 여부에 대한 평판 경향을 살펴보았다. 이 때 각 요인들이 긍정적 평판과 부정적 평판 평가 과정에 미치는 영향을 이해하기 위해서는 각 상황에서 발생하는 참가자들의 감정에 주목할 필요가 있다.

왜냐하면 감정은 생물학적 적합도를 최대화하기 위한 행동을 발생시키는 진화된 메커니즘으로써 상대방의 행동을 예측할 때 가장 중요한 단서로 사용되고, 어떤 행동에 대한 주관적 중요성을 바꿀 뿐 아니라(Fessler & Haley, 2003) 행동의 강도를 조절하는 기능을 가지고 있기 때문이다(Cosmides & Tooby, 2000). 실제 인간은 도덕적 행동이 요구되는 상황에 처하게 될 경우 특정 정서를 가지고 행동을 한 후 정서를 통해 행동을 합리화 하려는 경향이 있다(강진영, 1996). 이런 측면에서 감정은 참가자들이 상대방을 평판하는 과정을 이해하기 위한 중요한 단서이다.

긍정적 평판과 보상의 동기가 비슷하다고 가정할 때 참가자들이 긍정적으로 평판할 때 느끼는 주된 감정은 도덕적 찬동이자 감사의 정서이다. 도덕적 찬동은 공공의 이익을 위해 협력한 사람에게 상을 주고 싶어하는 정서이며, 감사는 상대방을 가치 있는 짝으로 여기고 기억되었음을 알림으로써 미래 상호작용에서 협력을 도모하는 방법이다(Fessler & Haley, 2003). 실험에서 긍정적 평판은 자신과 상대방 모두 협동하는 경우, 게임이 성공하는 경우, 그리고 상대방이 자신보다 자원이 적거나 자신과 같은 상황에서 협동하는 경우 더 강하게 나타났다. 각각의 상황에서 발생할 가능성이 있는 감정들을 살펴보면 다음과 같다.

시간, 자원, 노력, 기회비용을 투자하는 위험감수 행위는 집단 내 구성원간 신뢰와 연대감을 증가시킨다(Sosis, 2007). 실험 과정에서도 협동한 참가자는 다른 협동한 참가자와 함께 게임 실패에 의한 잠재적 위험을 감수하였을 뿐 아니라, 무임승차의 유혹을 뿌리침으로써 강한 연대감이 발생한 것으로 추정된다. 또한 평판 경쟁의 측면에서 협력을 선택한 참가자는 자신이 긍정적 평판을 받을 것을 예상할 수 있다. 따라서 어느 정도 평판 경쟁에서 자유로워진 상태에서 평판 점수를 부여하기 때문에 자신이 배신하였을 때보다 보다 긍정적으로 평판하였을 것이다. 한편 자신이 배신하였을 경우에 상대방이 협동하였을지라도 상대방에게 준 긍정적 점수는 상대방이 배신했을 때 준 긍정적 점수와 크게 차이가 나지 않았는데, 이 또한 자신보다 좋은 참가자 집단을 만나 유리해지는 것을 경계하려는 평판 경쟁으로 이해할 수 있다.

참가자들은 게임이 성공하고 상대방이 협동하였을 때 가장 긍정적으로 평판하였으며 그 다음 순위로 게임이 성공할 때 상대방이 배신을 하더라도 긍정적인 평판을 많이 하였다. 반면 게임이 실패한 경우 상대방의 협동 여부와 상관없이 긍정적 평판은 매우 약하게 나타났다. 게임에 성공한 참가자는 수익을 얻었고 이에 따른 연대감과 성취감을 통해 긍정적 평판을 내리는 경향이 나타났을 것이다. 이 때 특이한 점은 게임이 성공하면 상대방이 배신하였을지라도 긍정적으로 평판하는 눈감아주기 경향이 나타났는데 이러한 결과를 통해 상호작용의 과정에 따라서만이 아니라 상호작용의 결과에 의해서도 평판이 좌우될 수 있다는 것을 알 수 있다.

상대방의 자원이 자신보다 적거나 자신과 같은 상황에서 상대방의 자원이 자신보다 많은 상황보다 협동에 대한 긍정적 평판이 강하게 나타났다. 신호 비용의 하락은 신뢰성의 감소로 이어지며 이는 곧 그 신호 가치의 평가절하로 나타난다(Smith & Bird 2005; Zahavi & Zahavi, 1997). 자원을 많이 가진 사람이 협동을 하는데 드는 비용은 자원을 적게 가진 사람이 협동을 하기 위해 드는 비용보다 각각이 가지고 있는 자원에서의 비율로 볼 때 적다. 따라서 본 연구에서는 절대적 비용의 감소뿐 아니라 상대적인 비용의 감소 또한 행위 가치의 절하에 영향을 미칠 수 있음을 밝혔다. 한편 상대방이 배신할 경우 긍정적 평판은 절대적 양이 적어 자원의 상대적 차이에 따른 유의미한 차이가 나타나지 않았으나, 상대방의 자원이 자신보다 적은 경우, 상대방과 자신의 자원이 같은 경우, 상대방의 자원이 자신보다 많은 경우 순으로 참가자들이 긍정적 평판을 많이 부여하는 결과를 통해 부당한 평판하기 또한 부정적으로 편향되어 있음을 알 수 있다.

특이한 점은 자원의 불평등이 심화된 상황에서 자신과 자원의 양이 같은 참가자가 협동할 경우 가장 긍정적으로 평가하였다는 점이다. 이는 자신의 상황과 비슷한 타인에 대한 동질감을 기반한 공감 때문이라 생각한다. 공감은 상대방의 정서를 인식하고 타인의 감정을 대리적으로 경험하는 것으로서 타인에 대한 우호적 정서와 관심을 바탕으로 타인의 관점과 입장을 고려하는 태도이다(Hoffman, 2000). 공감은 타인의

정서 상태를 정확하게 판단함으로써 대인상황에 대처할 수 있는 수단을 강구할 수 있게 해주며 타인과 심리적 거리를 좁힐 수 있는 기술이 되기 때문에 (김현경, 2005) 개인에게도 진화적 이점을 갖는다. 이 때 동질감은 공감을 느끼기 위한 기본 요소 중 하나이며 익명의 실험에서 자원의 비슷한 상태는 참가자들 간 유일한 공통점으로 느껴질 수 있었을 것이다.

사람들이 부정적 평판을 할 때와 이타적 처벌을 할 때 느끼는 정서적 동기는 비슷하다(Feinberg, et al., 2012a). 따라서 부정적 평판의 주된 감정은 분노와 우월감을 느끼고 싶은 욕구일 것이다. 분노는 자신에게 피해를 주는 사람을 공격하여 미래에 피해가 반복될 가능성을 감소시키고, 무임승차자는 분노에 따른 보복을 피하고자 함으로써 협동이 촉진된다(Fessler & Haley, 2003). 우월감을 느끼고 싶은 욕구 또한 공격성으로 표출된다. 진화적 메커니즘에 따르면 타인보다 우위에 서고 싶은 욕구는 자신의 적합도를 상승시키려는 경쟁에서 승리를 위한 필수 요소이기 때문이다.

자신이 배신했을 때 상대방의 협동 여부와 상관없이 자신이 협동했을 때보다 부정적인 평판을 더 강하게 하였다. 배신한 사람은 타인이 협력할 가능성을 낮게 평가하였기 때문에 배신한 것이다. 따라서 이미 한 라운드를 진행하는 타인들에게 부정적인 선입견을 가지고 있기 때문에 부정적인 평판이 더 크게 나타났을 수 있다. 추가적으로 배신한 사람은 자신이 나쁜 평판을 받을 것을 이미 알고 있다. 따라서 평판에 있어서 자신이 상대적 우위를 점하기 위해서는 전략적으로 상대방의 협동 여부와 상관없이 부정적 평판을 줄 수 밖에 없는 상황에 처하게 된다. 따라서 배신한 참가자는 협력한 참가자보다 상대방에게 부정적으로 평판하였을 것이다.

게임의 성공 여부에 따른 부정적 평판은 게임이 실패한 경우에서 게임이 성공한 경우보다 강하게 나타났다. 심지어 게임이 실패하였을 때 배신한 참가자보다 협동한 참가자들이 부정적 평판을 많이 받았는데 이러한 결과는 참가자들이 상호작용의 과정보다 결과에 더 민감하게 반응하여 평판함을 나타낸다.

자원의 불평등이 심화된 상태에서 참가자들은 자신을 기준으로 자신보다 많은 자원을 가진 사람에 대해 자신보다 자원이 적거나 같은 사람에 비해 부정적 평판을 강하게 한다. 이 때 배신 행동에 대한 정당한 처벌로써 부정적 평판의 경우, 통계적으로 유의미하지는 않았으나 참가자들은 자신보다 자원이 많은 상대가 이기적으로 행동할 경우 가장 부정적으로 평판하였다. 이는 우위 개체에게 더 높은 협력의 기준이 적용되는 노블리스 오블리제 효과(Fiddick & Cummins, 2007)에 따라 동일한 수준의 기만에 대해 열등 개체가 더 강하게 처벌함을 증명했던 기존의 연구(김상인, 2006)와 일치하였다. 한편 협동했음에도 불구하고 부정적 평판을 내리는 부당한 평판의 경우에서 참가자들은 자원이 많은 상대방에게 가장 부정적으로 평판하였으며 통계적으로 유의미한 차이를 보였다. 결국 자원을 많이 가진 상태는 참가자들이 정당한 처벌로서 부정적 평판뿐 아니라 부당한 부정적 평판도 많이 받게 만듦으로써 부정적 평판 평가 편향을 심화시키며, 상대적 자원 차이에 따른 부정적 평판 평가 편향의 강화는 주로 부당한 평판에 의해 심화되었음을 알 수 있었다.

이러한 과정에서 질투의 감정이 크게 작동하였음을 예상할 수 있다. 질투란 자신 보다 나은 상대방과 자신의 차이를 줄이려는 감정을 말한다(Van de Ven et al., 2009). 이러한 경향은 평판이 처벌이나 보상에 비하여 비용 없이 질투란 감정을 표출할 수 있는 효과적이고 효율적인 수단이라는 이유 때문에 더욱 강화되었을 것이라 생각한다. 따라서 구성원간 불평등이 심할 때 구성원간 가진 자원이 비슷할 때 보다 질투의 감정이 크게 작동하였을 것이며 참가자간 평판을 내리는 과정에 영향을 주었을 것이다. 즉 참가자들은 자신보다 자원이 많은 상대에게 부정적 평판을 가해 다음 라운드에 평판이 좋지 않은 팀원과 만나게 만들어 자신과의 자원 차이를 줄이려 노력한 것이라 생각할 수 있다.

긍정적 평판과 마찬가지로 부정적 평판 또한 자신과 자원 수준이 같은 상대방에게 가장 자비롭게 평판하는 특이한 경향이 나타났다. 이는 동정에 의한 것이라 생각할 수 있다. 동정이란 괴로워하는 자의 처지에서 그들의 고통이나 어려움을 덜어주고자 하는

감정이다. 공감의 정서와 비슷하나 동정은 타인들에게 기쁨을 주는 데서 얻는 만족감과는 달리 타인들의 고통과 손해에 초점을 두고 있다(Blum, 1980). 자신과 비슷한 자원을 가진 상태는 더 쉽게 공감하고 동정이 발휘될 조건을 만들었을 것이다. 이로 인해 참가자들은 자신과 자원 수준이 같은 상대방에게 가장 덜 부정적으로 평판하는 경향을 보였을 것으로 추정된다.

Ⅷ. 연구의 한계 및 발전 방향

<실험1>보다 <실험2>에서 참가자들의 평판 평가가 더 공격적인 경향을 보였다. 소유한 자원의 불평등이 심화되었기 때문에 나타나는 현상이라 볼 수 있으나 실험의 수단이 도토리에서 현금으로 바뀌었기 때문에 가지는 심리적 효과와 실험이 순서대로 진행되었기 때문에 가지는 학습효과의 가능성도 무시할 수는 없다.

연구 결과 부당한 평판 부여가 많이 관찰되었는데 이는 평판을 하는데 따르는 위험요소를 무시하였기 때문이다. 특히 실제 사회에서 부당한 험담을 하였을 경우 신뢰도에 큰 타격을 받음으로써 보복을 당하거나 미래의 상호작용에서 제외 당함으로써 손해를 입을 수 있다. 그러나 이 연구에서는 부당한 평판을 제어할 수단이 없었다는 한계가 있다.

평판을 하는 행동도 보상이나 처벌에 비해 적지만 일정한 노력, 시간 등의 비용이 요구된다. 특히 부정적 평판을 퍼뜨리는 사람의 경우 자신의 평판이 훼손될 수 있다는 위험요소가 있다. 실제 무임승차자를 처벌하는 사람은 집단 지향적이고 신뢰할만하지만 좋은 사람이라 인식되진 않는다(Barclay, 2006). 또한 험담의 대상으로부터 보복 당할 위험도 무시할 수는 없다. 따라서 평판 점수를 부여하는데 긍정적 평판 1점에 10원, 부정적 평판 1점에 30원과 같이 적은 액수를 설정하였다면 부당한 평판을 줄이면서 더욱 정확한 실험이 될 수 있었을 것이다.

한편 실험에서 참가자들은 추가적인 이득이 없었음에도 불구하고 평판 평가에 매우 적극적으로 참여하였다. 그러나 평판을 하는 행동 자체가 평판하는 사람에게 주는 이득 또한 무시할 수 없다. 타인의 평판 정보를 전달하는 행동은 행위자가 이미 무임승차자에 대한 평판을 퍼뜨림으로써 무임승차자의 또 다른 이기적 행동을 방지할 수 있다는데 직접적인 이득이 있고, 행위자의 넓은 사회적 인맥을 과시함으로써 지위를 높일 수

있으며, 그가 집단지향적이고 믿을만한 사람이라는 것을 주위사람들에게 알릴 수 있다는 이점이 있다(Willer, et al., 2010). 따라서 향후 연구에서는 참가자가 정당한 평판을 한 경우 적지만 추가적인 이익을 제공할 필요가 있다.

마지막으로 평판 점수가 10점으로 제한되어 있었는데 점수의 총 양이 적다고 느끼는 참가자가 많았다. 특히 여러 명이 배신을 하였을 때 10점을 분배하여 나눠야 하기 때문에 부정적 평판을 부여하는데 한계가 있을 수 있었다. 따라서 부여할 수 있는 평판 점수의 총 양을 좀더 늘릴 필요성이 있다.

Ⅸ. 참고 문헌

Alexander, R. (1986) Ostracism and indirect reciprocity: The reproductive significance of humor, *Ethology and sociobiology*, 7, 253-270.

Axelrod, R. (1984) *The evolution of cooperation*. New York. Basic Books.

Bateson, M., Nettle, D., & Roberts, G. (2006) Cues of being watched enhance cooperation in a real-world setting, *Biology Letters*, 22; 2(3): 412-414.

Barclay, P. (2006) Reputational benefits for altruistic punishment, *Evolution and Human Behavior*, 27, 325-344.

Barclay, P., & Willer, R. (2007) Partner choice creates competitive altruism in human. *Proceedings of the royal society of london*, Series B: Biological Sciences, 274, 749-753.

Bereczkei, T., Birkas, B., Kerekes, Z. (2007) Public charity offer as a proximate factor of evolved reputation-building strategy: an experimental analysis of a real-life situation. *Evolution and Human Behavior*, 28, 277-284.

Bereczkei, T., Birkas, B., Kerekes, Z. (2010) Altruism towards strangers in need: costly signaling in an industrial society. *Evolution and Human Behavior*, 31, 95-103.

Bird, B., Rebecca, L., Douglas, W., Smith, A. (2002) Risk and reciprocity in Meriam food sharing. *Evolution and Human Behavior,* 23(4), 297-321.

Blum, L. (1980) Compassion. In A. O. Rorty (Eds.) *Explaing emotions*, 507-517. Berkeley: university of California press.

Boone, J. L. (1998). The evolution of magnanimity: When is it better to give than to receive? *Human Nature*, 9, 1-21.

Boyd, R., & Richerson, P, J. (1989) The Evolution of Indirect Reciprocity, *Social Networks*, 11, 213-236.

Boyd, R., Gintis, H., Bowles, S., Richerson, J. P. (2005) The Evolution of Altruistic Punishment. In Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (Eds.) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life,* 215-227. Cambridge, MIT Press.

Chiappe, D., & Brown, A. (2004) Cheaters are looked at longer and remembered better than cooperators in social exchange situations, *Evolutionary Psychology,* 2, 108- 120.

Cosmides, L., & Tooby, J. (1992) Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture,* 163-228. New York7 Oxford University Press.

Cosmides, L., & Tooby, J. (2000) Evolutionary Psychology and the Emotions In M. Lewis & J. M. Haviland-Jones. (Eds.) *Handbook of emotion*, 91-115. New York, Guilford Press.

Dunbar, R. (2004). Gossip in evolutionary perspective. *Review of General Psychology,* 8(2), 100-110.

Fehr, E., & Fischbacher, U. (2004) Third-Party Punishment and Social Norms. *Evolution and Human behavior*, 25(2), 63-87.

Fehr, E., & Gächter, S. (2002) Altruistic Punishment in Humans, *Nature*, 415, 137-140.

Feinberg, M., Cheng, J. T., & Willer, R. (2012). Gossip as an effective and low cost form of sanctioning. *Brain and Behavioral Sciences*, 35(1), 25.

Feinberg, M., Willer, R., Stellar, J., Keltner, D. (2012) The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*, 102(5), 1015-1030.

Fessler, M.D. & Haley, J.K. (2003) The Strategy of Affect: Emotions in Human Cooperation. In Hammerstein, P. (Eds.) *Genetic and Cultural Evolution of Cooperation*, 7-36. Cambridge, MA: MIT Press.

Fiddick, L. & Cummins, D. D. (2007) Are Perceptions of Fairness Relationship-Specific? The Case of Noblesse Oblige. *Quarterly Journalof Experimental Psychology*, 60, 16-31.

Fox, E., Russo, R., & Dutton, K. (2002) Attentional bias for threat: Evidence for delayed disengagement from emotional faces. *Cognition and Emotion*, 16, 355-379.

Gürerk, Ö., Irlenbusch, B., Rockenbach, B. (2004) On the Evolvement of Institutions in Social Dilemmas.

Haley, K. J. & Fessler, D. M. T. (2005) Nobody' s watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*. 26, 245-256.

Hamilton, W. (1964) The genetical evolution of social behaviour. I, *Journal of Theoretical Biology*, 1-16.

Hardy, C, L., Van Vugut, M. (2006) Nice Guys Finish First: The Competitive Altruism Hypothesis. *Personality and Social Psychology Bulletin*, 32(10), 1402-1413.

Hawkes, K. (1992). Sharing and Collective Action. In Smith, E. A., Winterhalder, B., Hawthorne, (Eds.) *Evolutionary Ecology and Human Behavior*. 269-300. New York: Aldine de Gruyter.

Hawkes, K. (1993). Why Hunter-Gatherers Work. *Current Anthropology* , 34(4), 341-362.

Hoffman ,L. M. (2002) *Empathy and moral development : Implications for caring and justice*. NewYork: Cambridge press.

Hoffman, E., McCabe, K., Shachat, K., Smith, V. (1994) Preferences, Property Rights, and Anonymity in Bargaining Games, *Games and Economic Behavior*, 7, 346-380.

Ito, T, A., Larsen, J, T., Smith, N. K., Cacioppo, J, T. (1998) Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, 75(4), 887-900.

Kaplan, H., Hill, K. (1985). Food Sharing among Ache Foragers: Tests of Explanatory Hypotheses. *Current Anthropology*, 26, 223-233.

Marlowe, W., Berbesque, J. C.,Barr, A.,Barrett, C.,Bolyanatz, A., Cardenas, C.J., Ensminger,J., Gurven, M., Gwako,E., Henrich,J., Henrich,N., Lesorogol,C., McElreath, R., &Tracer, D. (2008) More "altruistic" punishment in larger societies. *Proceedings of Royal Society of London*, 275B, 587-590.

Milinski, M., Semmann, D., Bakker, T. C., Krambeck, H. J. (2001) Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proceedings of the Royal society of London*, Series B: Biological sciences, 268, 2495-2501.

Milinski, M., Semmann, D., & Krambeck, H. (2002) Donors to charity gain in both indirect reciprocity and political reputation, *Proceeding of The Royal Society,* 269, 881-883.

Milinski, M., Semmann, D., & Krambeck, H. (2005) Reputation is valuable within and outside one' s own social group, *Behavioral Ecology and Sociobiology*, 57(6), 611-616.

Nowak, M. A., Sigmund, K. (1998) Evolution of Indirect Reciprocity by Image Scoring, *Nature,* 393, 573-577.

O¨hman, A., & Mineka, S. (2001) Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, 108, 483 - 522.

Oda, R., Niwa, Y., Honma, A., Hiraishi, K. (2011). An eye-like painting enhance the expectation of a good reputation. *Education and Human Behavior,* 32, 166-171.

Ohtsuki, H., & Iwasa, Y. (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature,* 457, 79-82.

Panchanathan, K., & Boyd, R. (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432, 499-502.

Peetersa, G. & Czapinskib, J. (1990) Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects. *European Review of Social Psychology*, 1(1), 33-60.

Piazza, J., & Bering, J. M., (2008). Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior*, 29, 172-178.

Pollock,G., & Dugatkin, L. A. (1992) Reciprocity and the emergence of reputation, *Journal of Theoretical Biology*, 159, 25-37.

Price, M,E., Cosmides, L., Tooby, J. (2002) Punitive sentiment as an anti- free rider psychological device, *Evolution and Human Behavior*, 23, 203-231.

Rigdon, M., Ishii,K., Watabe, M.,& Kitayama, S. (2009) Minimal Social Cues in the Dictator Game, *Journal of Economic Psychology*, 30(3), 358-367.

Rockenbach, B., & Milinski M. (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444, 718-723.

Roth, A,E., Vesna, P., Masahiro, O., Shmuel, Z. (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an Experimental study, *American economic review*, 81, 1068-1095.

Sachs, J,L., Mueller, U,G., Wilcox, T,P., & Bull, J,J. (2004) The evolution of cooperation. *The Quarterly Review of Biology,* 79, 135-160.

Smith, E,A., Bird, R,L,B. (2005) Turtle Hunting and Tombstone Opening: Public Generosity as Costly Signaling. *Evolution and Human Behavior,* 21(4): 245-261.

Sommerfeld R. D., Krambeck, H., Semmann, D., & Milinski, M. (2007) Gossip as an alternative for direct observation in games of indirect reciprocity, *Proceeding of the National Academy of Science* 104, 17435-17440.

Sosis, R., Kress, H, C., Boster, J, S. (2007) Scars for war: evaluating alternative signaling explanations for cross-cultural variance in ritual costs Original Research Article, *Evolution and Human Behavior*, 28(4), 234-247.

Taylor, S, E. (1991) Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, 110(1), 67-85.

Trivers, R, L. (1971) The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46(1): 35-57.

Van de Ven, N., Zeelenberg, M., Pieters, R. (2009) Leveling up and down: The experiences of benign and malicious envy. *Emotion,* 9, 419-429.

Vanneste, S., Verplaetse, J., Hiel, A. V., Braeckman, J. (2007) Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior*, 28, 272-276.

Wedekind, C., & Milinski, M. (2000) Cooperation of indirect reciprocity by image scoring. *Nature*, 393, 573-577.

Wedekind, C.,& Braithwaite, V. A., (2002) The long term benefits of human generosity in indirect reciprocity. *Current biology*, 12, 1012-1015.

Willer, R., Feinberg, M., Irwin, K., Schultz, M., & Simpson, B. (2010) The Trouble with Invisible Men: How Reputational Concerns Motivate Generosity. In Hitlin, S. &Vaisey, S. (Eds.) *The Handbook of the Sociology of Morality.* 315-330. New York: Springer.

Wilson, M & Daly, M (1985) Competitiveness, risk taking, and violence: the young male syndrome Original Research Article. *Ethology and Sociobiology*, 6(1), 59-73.

Zahavi, A. (1975) Mate selection- a selection for a handicap. *Jounal of theorical biology*, 53, 205-214.

Zahavi, A., & Zahavi, A. (1997) *The handicap principle: A missing piece of Darwin's puzzle*. New York: Oxford University Press.

강진영 (1996) 도덕 교육에 있어서 정서의 역할에 관한 연구. 건국대학교 대학원 박사학위논문

김상인 (2006) 구성원 간 불평등이 집단 내 협력에 미치는 영향: 공공재 게임을 통한 진화심리학적 연구. 서울대학교 대학원 인류학과 석사학위논문

김현경 (2005) 도덕적 행동요인으로서 정서의 도덕교육적 함의 연구. 서울대학교 대학원 국민윤리교육과 석사학위논문

통계청 (2012) 2012 년 2/4 분기 가계동향

<Abstract>

Characteristics of reputation evaluation in the public goods game

Seewon Lee
Department of Anthropology
Graduate School
Seoul National University

The altruistic behavior of humans is interesting in that it is not limited to kin and that it takes place despite the uncertainty of the future interaction with the recipient. Based on the fact that there is a difference in altruistic behavior depending on the existence of an observer, one can infer that altruistic behavior is a form of insurance for future interactions. People show altruistic behavior to enhance good reputation and avoid negative reputation. This is because reputation has the effect of both compensation and punishment. A good reputation, as compensation, increases the chances of receiving help or being preferred as a mate or partner in future interactions. On the other hand, a negative reputation, as punishment, increases the chance of social alienation. In other words, reputation has the characteristics of deferred compensation and punishment. The critical factor of reputation is that the cost of reputation is substantially lower than compensation and punishment, and that information is easily shared between members. Therefore reputation is an effective factor that sustains altruistic cooperation in a society and considering reputation serves an important role in supplementing the limitations in prior literature on altruistic behavior.

Despite such advantages, prior literature does not adequately reflect the characteristics of reputation before including it in their experiments. Prior literature has overlooked the following characteristics of cooperation: First, prior literature includes reputation as a score that is set by the researcher instead of as the actual assessment of the participants. Second, the level of increase from good reputation is assumed to be the same with the decrease from negative reputation, which is an oversimplification. Third, the reputation evaluation on others could be different depending on whether the participant cooperated or whether the game succeeded. However this factor is frequently omitted from the research design as well. Finally, prior literature has imposed the same reputation system on every participant without considering the unequal distribution of resources.

Therefore, in order to consider the characteristics of reputation, this thesis constructs a public goods game that distinguishes positive reputation and negative reputation. The reputation is measured by having participants evaluate each other's reputation. Through this research design, the results show that in the evaluation, the negative reputation from selfish behavior is stronger than the positive reputation from altruistic behavior in the evaluation of others, the negative reputation from selfish behavior dominates the positive reputation from altruistic behavior.

The research also reveals that reputation is influenced by not only the cooperation of the counterpart, but also the participant's cooperation, success of the game, and the relative difference in resource distribution. In order to examine the influence of the participant's cooperation in the reputation evaluation of the counterpart's cooperation, the research sets 4 situations based on the cooperation

of the participant and the counterpart: cooperation–cooperation, cooperation–betrayal, betrayal–cooperation, and betrayal–betrayal. The positive and negative reputation on counterparts were measured and compared by situation. The participants evaluated their counterparts more positively only in the cooperation–cooperation situation. On the other hand, they evaluated their counterparts more negatively when they betrayed compared to when they cooperated, regardless of the counterpart's cooperation or betrayal.

In addition, this study examines how the result of the game influences the reputation evaluation on the counterpart's reputation. Based on the success of the game and the counterpart's cooperation, the research examines and compares the reputation evaluation in 4 situations: success–cooperation, success–betrayal, failure–cooperation, and failure–betrayal. The positive reputation evaluation was stronger when the game was successful. In particular, positive evaluation was the strongest when the game was successful and the counterpart cooperated. If the game was successful, the positive reputation evaluation was stronger even if the counterpart betrayed. On the other hand, participants gave harsher negative reputation evaluations when the public goods game had failed compared to when the game was successful, regardless of the counterpart's cooperation.

By distributing different amounts of monetary resources in the public goods game with identical rules, how reputation is influenced by resource inequity was examined. The counterpart with relatively more resources was less likely to receive a positive evaluation even if the counterpart cooperated, and was more prone to negative reputation evaluation. In addition, against the expectation that negative reputation evaluation would be weakest if the counterpart had relatively less resources, participants were likely to give more positive reputation

evaluations and less negative reputation evaluations to counterparts that had a similar amount of resources.

Finally, the extent of the influence of each factor on reputation evaluation was examined. Positive reputation evaluation was strongest in the order of similar resources, successful results of game, participant's cooperation, and counterpart's cooperation. As for negative reputation evaluation, the influence was stronger in the order of counterpart with more resources, participant's betrayal, failure of the game, and counterpart's betrayal.

Keywords: Altruistic behavior; Public goods game; Reputation evaluation; Resource inequity

Student Number: 2010-22986

<감사의 말>

무엇보다 먼저 박순영 지도 교수님께 가장 큰 감사의 말씀을 드립니다. 교수님과 함께 했기에 서울대학교에서의 모든 과정을 기쁘게 마무리 할 수 있었습니다. 박순영 교수님께서는 제가 학부 시절부터 진화심리학에 관심을 가지고 보다 넓은 학문적 경험을 할 수 있도록 도와주셨습니다. 또한 제가 방황하고 고민할 때 교수님의 따뜻한 배려 덕분에 용기를 잃지 않고 여유를 찾아 논문을 진행할 수 있었습니다. 논문의 미숙한 진행에도 불구하고 애정이 담긴 격려와 충고로 지도해주셔서 정말 감사합니다. 교수님의 냉철하고 명료한 가르침과 지혜는 대학원 공부에 있어서뿐만 아니라 저의 가치관의 형성에도 큰 도움이 되었습니다. 교수님께 존경을 표하며 다시 한번 감사 드립니다.

다음으로 논문의 심사에 참여해주신 이현정 교수님과 전중환 교수님께 감사 드립니다. 교수님들의 칭찬과 응원이 논문을 끝까지 완성할 수 있는 힘이 되었습니다. 또한 실험과 모의실험에 참여해준 서울대학교 학우들, 게임의 설계에 도움을 준 정재문군, 통계 분석에 도움을 준 백승학군, 실험의 진행에 도움을 준 김용균군에게도 감사의 마음을 전합니다.

마지막으로 논문의 퇴고를 부탁했으나 귀찮다는 핑계로 거절한 가족들에게 평온한 보금자리라도 제공해줘서 고맙다는 말을 전합니다.